It's the thought that counts:
The role of intentions in noisy repeated games

First version: April 13 2013
This version: May 5, 2015

David G. Rand[a], Drew Fudenberg[b] and Anna Dreber[c]

[a]Corresponding author. Department of Psychology, Department of Economics, Cognitive Science Program, School of Management, Yale University, Box 208205, New Haven CT 06520-8205, e-mail: david.rand@yale.edu, phone: +12034324500**.**
[b]Department of Economics, Harvard University
[c]Department of Economics, Stockholm School of Economics
Rand and Fudenberg are co-lead authors.

Abstract:

We examine cooperation in repeated interactions where intended actions are implemented with noise but intentions are perfectly observable. Observable intentions lead to more cooperation compared to control games where intentions are unobserved, allowing subjects to reach similar cooperation levels as in games without noise. Most subjects condition exclusively on intentions, and use simpler, lower-memory strategies compared to games where intentions are unobservable. When the returns to cooperation are high, some subjects are tolerant, using good outcomes to forgive attempted defections; when the returns to cooperation are low, some subjects are punitive, using bad outcomes to punish accidental defections.

Keywords: cooperation, prisoner's dilemma, repeated games, intentions
JEL codes: C7, C9, D00

1. Introduction

This paper studies cooperation in infinitely repeated games where the intended actions are implemented with error, so that the actions played are only a noisy or implicit signal of what was intended. The possibility of error is pervasive in social interactions, and many if not most of these interactions do not have a fixed and known termination date. The resulting imperfect public monitoring has received a large amount of attention in the theoretical literature on infinitely repeated games (e.g., Green and Porter 1984, Radner et al. 1986, Abreu et al. 1990, Fudenberg et al. 1994), but only a handful of experimental studies have explored infinitely repeated games with errors (e.g. Aoyagi and Frechette 2009, Bigoni et al. 2012, Fudenberg et al. 2012, Aoyagi et al. 2013).[1]

Our setup differs from that of these past studies in that we consider the effect of players directly observing the intended actions of their opponents, in addition to the realized ones. This sort of information is available in some real-world settings, for example compensation for hedge fund managers where both the positions taken and the actual outcomes are observable and thus explicit, or in a homicide when it is clear that the accused shot the victim but extenuating circumstances may exist – here the legal system pays attention to both intentions and outcomes, differentiating between manslaughter and various levels of murder.

From a theoretical standpoint, the impact of explicitly observing intentions is clear: the highest equilibrium payoff can be obtained with strategies that completely ignore the realized outcomes and condition only on intended play, and moreover this best equilibrium is the same as when actions are implemented without error. Note that this is very different from the situation in one-shot games, where maximizing monetary payoffs would lead subjects to ignore intentions entirely. Even in those games, a substantial proportion of subjects do condition on intentions in addition to outcomes when both pieces of information are available.[2] One possible explanation for this apparent "preference for reciprocity" is that it reflects a heuristic that fosters cooperation in repeated interactions. If so, we might expect to see even more reliance on intentions in settings where conditioning on intentions leads to a cooperative equilibrium even in the absence of a

[1] Van Lange et al. (2002) studied play in a repeated continuous-choice (rather than binary) PD with errors where subjects were matched against computer partners playing either tit-for-tat or tit-for-two-tats (but were told that the partners were actual people).

[2] Past work on intentions in one-shot games is discussed in Section 2, as well as the Bereby-Meyer and Roth (2006) and Kunreuther et al. (2009) studies of intentions in the finitely-repeated prisoner's dilemma.

preference for reciprocity. At an empirical level, the question of how extensively people condition their play on intentions in infinitely repeated games remains open, as does the extent to which they also condition on outcomes, and the effect of all this on the level of cooperation.

To begin to understand these issues, we study the experimental play of the repeated prisoner's dilemma when intended actions are implemented with error. Our main goals are to understand when and in what ways subjects use data on intentions and outcomes, and how cooperation when intentions are revealed compares to either a setting with error when intentions are not observed, or one in which error is not present (so the actions themselves reveal the intentions). Our primary experiment presents evidence from a set of infinitely repeated prisoner's dilemma games with a continuation probability of 7/8 and an error rate of 1/8. In our main treatments, intentions are explicit; as controls, we also consider the same games but where only actions are observable (thus leaving intentions implicit), as well as the same games without exogenously imposed error (where the observed action corresponds to the intended one). We explore two different payoff specifications for the stage game actions "Cooperate" ("C") and "Defect" ("D") (neutral language was used in the experiment itself). In the "high benefit" treatment, the benefit that playing "C" gives to the other player is high enough that there is a cooperative equilibrium in the game with errors whether or not intentions are observed. In the "low benefit" treatment, the benefit that C gives is low enough that the only equilibrium with errors and unobserved intentions is for both players to always defect, although cooperation remains an equilibrium outcome when intentions are observed.

Summary of results

We use two different methods to analyze the PD data: a structural estimation of the distribution of strategies using the "structural frequency estimation method" (SFEM) of Dal Bó and Frechette (2011), and a descriptive analysis that relates play in a given period of a supergame to the opponent's intention and action in the period before (which implicitly assumes subjects use strategies that mostly depend on that information). Both methods show that most subjects condition almost exclusively on intentions and thus play consistently with predictions based on maximizing money payoffs: In our descriptive analysis, the effect of opponent's intention is dramatically larger than that of the actual outcome. Similarly, in the strategy estimation, more than two thirds of subjects use strategies that do not condition on outcomes.

To the extent that subjects do condition on outcomes, interestingly, they do so in different ways depending on the payoff specification. In the treatment where there is less of an incentive to cooperate, some people (about 15%) are punitive in treating both accidental cooperation (partner meant to play D but played C) and accidental defection (partner meant to play C but played D) as defection; only when the partner both intended to play C and actually did so was this treated as cooperation. This behavior is not observed in the treatment with high returns to cooperation, where instead some people (about 19%) are tolerant in that they only retaliate against intentional defections – these subjects forgive both accidental defection (partner meant to play C but played D) as well as accidental cooperation (partner meant to played D but played C). Thus the "punitive" subjects in the low-benefit treatment use realized outcomes to punish cooperators that defect by accident, while in the high-benefit treatment "tolerant" subjects use the realized outcomes to forgive defectors that accidentally cooperated.

By conditioning largely on intentions, subjects are able to achieve high levels of cooperation in both treatments. Compared with the controls in which intentions are implicit, explicitly revealing intentions lead to significantly more cooperation. Interestingly, this increase in cooperation is not associated with more leniency (where the subject overlooks the partner's first defection) but instead with an increase in simple strategies that conditioned on at most the previous period. This suggests that many of the longer memory strategies seen in Fudenberg et al. (2012) were the result of subjects trying to infer the intentions of their opponent, either because doing so leads to higher monetary payoffs or because preferences depend on the intentions of others.

In principle, games with errors but explicit intentions are distinct from games with no errors, so people might use different strategies in each. To evaluate this possibility, we compare play when intentions are explicit to play in games where there are no errors.[3] We find that not only does revealing intentions increase cooperation compared to games where intentions are implicit, but it successfully moves cooperation levels all the way back up to the level seen in the absence of errors. Revealing intentions also leads subjects to use similar strategies to those

[3] Previous work on infinitely repeated games without errors has shown that subjects learn to cooperate, as long as the returns on cooperation are large enough relative to the continuation probability (Dal Bó 2005, Dreber et al. 2008, Dal Bó and Frechette 2011, Fudenberg et al. 2012, Rand and Nowak 2013). Furthermore, cooperation is significantly higher without errors compared to the case with errors where intentions are implicit (Fudenberg et al. 2012).

observed in the absence of errors; in particular in both cases most subjects condition only on play in previous period.

2. Experimental Design and Empirical Questions

In our experiments, the infinitely repeated prisoner's dilemma was induced by having a known constant probability of $\delta = 7/8$ that a supergame would continue between two players following each period; with probability 1-δ, the supergame ended and subjects were informed that they have been re-matched with a new partner. In the main treatments, there was also a known constant error probability of E=1/8 that an intended move is changed to the opposite move. In this "explicit intentions" treatment, subjects were informed of the intended action of the other player, the other player's realized action, and whether their own move had been changed (i.e. when they make an error). We also have an "implicit intentions" control, where subjects were told their own realized action and the realized action of the other player but not the other player's intended action. Finally, we have a set of control conditions without errors on realized payoffs. Note that some of the control conditions (but not the explicit-observed intentions treatment) were reported in Fudenberg et al. (2012): we explain more on this later in the paper.

Realized payoffs **Expected payoffs**

b/c = 1.5

	C	D
C	1,1	-2,3
D	3,-2	0,0

b/c = 1.5, E = 1/8

	C	D
C	0.875, 0.875	-1.375, 2.375
D	2.375, -1.375	0.125, 0.125

b/c = 4

	C	D
C	6,6	-2,8
D	8,-2	0,0

b/c = 4, E = 1/8

	C	D
C	5.25, 5.25	-0.75, 6.75
D	6.75, -0.75	0.75, 0.75

Figure 1. *Payoff matrices for each condition. Payoffs are denoted in points.*

Subjects were informed of the specifics of their treatment (but not the existence of other treatments) in the experimental instructions, which are included in the Online Appendix.

The stage game is the prisoner's dilemma in Figure 1 where the payoffs are denoted in points. Cooperation and defection take the "benefit/cost" (b/c) form, where cooperation means paying a cost c to give a benefit b to the other player, while defection gives 0 to each party; b/c took the values 1.5 and 4.[4] The expected payoff tables in Figure 1 incorporate the noise probability E=1/8. Subjects were presented with both the b/c representation and the resulting pre-error payoff matrix, in neutral language (the choices were labeled A and B as opposed to the "C" and "D" that is standard in the prisoner's dilemma).[5]

As noted earlier, in the explicit-intentions treatment, the highest equilibrium payoffs can be supported with strategies that condition only on intentions and ignore outcomes; moreover, the set of such equilibria is the same as in a game with the same expected payoff matrix and explicit actions. Under both of the payoff specifications we used, the explicit-intentions game has subgame-perfect equilibria in which both players cooperate each period, including for example the strategy profile where both players use the "Grim" strategy, which is "Play C iff either player has never played D". However, Dal Bó (2005) shows that the existence of a cooperative equilibrium in a repeated game without noise is not sufficient for there to be much cooperation, and subsequent work by Blonski et al. (2011), Dal Bó and Frechette (2011), (2013) and Rand and Nowak (2013) suggests that a key determinant is whether Grim risk-dominates the strategy "Always Defect" (ALLD) in a 2x2 game. We might suspect that a similar pattern would apply to games with noise and observed intentions, so we note that "Grim-I" (the grim strategy that conditions only on intentions and ignores outcomes) risk-dominates ALLD even in the low-benefit treatment.[6]

Turning to the game with errors and implicit intentions, we note that in the low-benefit treatment b/c=1.5 "both play Grim" is not a Nash equilibrium but this is an equilibrium when

[4] Each session used a single payoff specification. Note that the benefit/cost specification implies that the short-run gain to playing D instead of C is independent of the other player's action. The prisoner's dilemma is more general than this; its defining characteristics are that D is a dominant strategy and that both playing C yields the highest payoff - in particular both playing C should be more efficient than alternating between (C,D) and (D,C).

[5] We use negative payoffs in order to illustrate that cooperation entails paying a cost for someone else to recieve a benefit, allowing us to break down C and D into costs and benefits for each player independently of what the other player does.

[6] In the noisy repeated game Grim-I earns a discounted average payoff of 7/8 when facing itself. Grim-I vs ALLD yields -11/8 the first period, then 1/8 afterwards for discounted average of - 4/64; ALLD vs Grim-I earns 26/64; and ALLD vs ALLD earns 1/8. Thus facing a 50-50 mixture between the two strategies, Grim-I earns 26/64, while ALLD gets 17/64.

b/c=4.[7] Consistent with this observation, the experiments reported in Fudenberg et al. (2012) found substantially more cooperation at b/c=4 than when b/c=1.5.

As noted above, some of the data in our control conditions were originally reported in Fudenberg et al. (2012): in particular, the implicit intentions condition for both b/c=1.5 and b/c=4 and the no-error condition for b/c=4 were originally reported there.[8] The new conditions for this paper are the explicit intentions conditions for both b/c=1.5 and b/c=4 and the no-error condition for b/c=1.5, summing up to a total of 6 new sessions with 128 participants. These sessions were conducted in 2013. All in all, including both the previously reported control conditions and these new sessions, a total of 338 subjects participated at the Harvard Decision Science Laboratory in Cambridge, MA. In each session, 12-32 subjects interacted anonymously via computer using the software z-Tree (Fischbacher 2007) in a sequence of indefinitely repeated prisoner's dilemmas (see Table 1 for summary statistics on the different conditions). In total, we conducted a total of 16 sessions between September 2009 and December 2012. We only implemented one condition during a given session, so each subject participated in only one condition. We used the exchange rate of 30 units = $1. Subjects were given a show-up fee of $10 plus their winnings from the repeated prisoner's dilemma.[9] To allow for negative stage-game payoffs, subjects began the session with an "endowment" of 50 units (in addition to the show–up fee).[10] On average subjects made $18 per session, with a range from $11 to $32. Sessions lasted approximately 60 minutes.[11]

[7] See the online appendix to Fudenberg et al. (2012) for equilibrium calculations of the implicit-intentions game. Note that there are many other cooperative equilibria in this game when b/c=4, including "perfect Tit for tat", which says to play C if yesterday's outcome was (C,C) or (D,D) and otherwise play D.

[8] All sessions were conducted during the academic year, and all subjects were recruited through the CLER lab at Harvard Business School using the same recruitment procedure. In particular, the implicit intentions treatments and the no-error treatment for b/c=4 were originally reported in Fudenberg et al. (2012). These earlier controls used the more demanding "turnpike protocol" as a way to rule out contagion effects. However as the turnpike protocol restricts the number of supergames to one-half of the subjects in the room, our subsequent work has replaced it with the more common random-matching protocol. The meta-study of Dal Bó and Frechette (2015) shows that the turnpike does not affect cooperation rates using almost 40,000 observations across studies (the coefficient estimate is both very small in magnitude and not statistically significant).

[9] Subjects also received earnings from a post-PD allocation decision that they were unaware of when playing the PD.

[10] No subject finished with an endowment of fewer than 19 units, and only 2 out of 338 subjects had fewer than 50 units.

[11] Subjects were given at most 30 seconds to make their decision, and informed that after 30 seconds a random choice would be made. The frequency of random decisions was very low, only 163 out of 28,664 decisions. Furthermore, only 16 out of 228 subjects ever ran out of time. The largest proportion of random choices for any individual subject was 0.124.

Table 1. *Summary statistics per condition and b/c.*

	b/c=1.5			b/c=4		
	Explicit intentions	No error	Implicit intentions*	Explicit intentions	No error*	Implicit intentions*
Sessions per condition	2	2	3	2	3	4
Subjects per condition	44	44	72	40	48	90
Average number of supergames	9.5	9	11	9.5	8	11.3
Average number of periods per supergame	8.2	8.2	8.4	8.1	8.2	8.1

*Note that these conditions were reported in Fudenberg et al. (2012).

To implement random game lengths, we followed the procedure of Dreber et al. (2008) and Fudenberg et al. (2012): In each session every first supergame lasted t_1 periods, every second supergame lasted t_2 etc. For comparability between the implicit and explicit intentions data, we used the sequence of game lengths generated in Fudenberg et al. (2012), completing as many games as possible within the allotted session time.[12]

Past experiments on revealed intentions in games with errors have only studied one-shot or finitely repeated games. Bereby-Meyer and Roth (2006) explore cooperation in the one-shot and finitely repeated prisoner's dilemma where actions are implemented without noise and payoffs are either a deterministic or stochastic function of the actions played; since the end period is common knowledge, there is a substantial last-period effect in each supergame (as in Kunreuther et al. 2009 who also explore random payoffs). They find that the outcome due to the random shock in the previous period matters for the decision to cooperate this period, but less so than whether the other player cooperated or defected in the previous period.[13] Charness and Levine (2007), Cushman et al. (2009), Schächtele et al. (2011) and Rubin and Sheremeta (forthcoming) study one-shot games where intentions and outcomes can be each be either good or bad, can be in conflict due to a random device, and can either be rewarded or punished.[14]

[12] Note that from the viewpoint of the subjects, it was irrelevant when the game lengths were determined. The length of each interaction was as follows: 3 round practice; 8, 7, 10, 7, 8, 9, 5, 11, 9, 8 ,7, 8. Note that some of the previously collected sessions from Fudenberg et al. (2012) deviated slightly from this order; see Fudenberg et al. (2012) for details.

[13] Aoyagi and Frechette (2009) and Ambrus and Greiner (2012) study repeated games with imperfect public monitoring; but subjects do not observe their opponents' intended actions.

[14] Additional work exploring the role of intentions in one-shot games is described in papers such as Blount (1995), Brandts and Solà (2001), Andreoni et al. (2002), Falk et al. (2003), McCabe et al. (2003) and Falk et al. (2008), which either compare how subjects respond to offers made by humans versus randomly generated offers (thus removing intentionality) or vary the strategy space of one player (thus changing the intentionality associated with a given outcome). In these papers intentions and outcomes are not in conflict.

While a significant fraction of subjects in these games at least partially condition on intentions, there is also a tendency for them to condition on outcomes. Because these settings do not have a cooperative equilibrium, this work offers little guidance as to the fraction of players that will follow the theoretically optimal policy of conditioning only on intentions in the explicit-intentions treatment, on how the players who do respond to outcomes will do so, or on how play in the explicit-intentions and no-error infinitely repeated games will compare. Note that although conditioning only on intentions yields the highest equilibrium payoffs, a subject who believes that other subjects will condition on outcomes as well as intentions will find it optimal to do so as well, as may a subject who is uncertain whether others respond to outcomes as well as intentions.

Inspired both by past experimental findings and theoretical concerns, we organize our analysis around the following questions:

QUESTION 1: Does observing intentions allow more cooperative play compared to no intentions?

QUESTION 2: How similar are cooperation rates in the explicit-intentions treatments compared to the no-error treatments?

QUESTION 3: How close do subjects come to basing their play solely on intentions?

QUESTION 4: To the extent that subjects condition on realized outcomes as well as intentions in the explicit-intentions treatment, how do they do this?

QUESTION 5: How do the strategies used in the explicit-intentions games compare to those used in games with implicit intentions or no errors?

3. Results

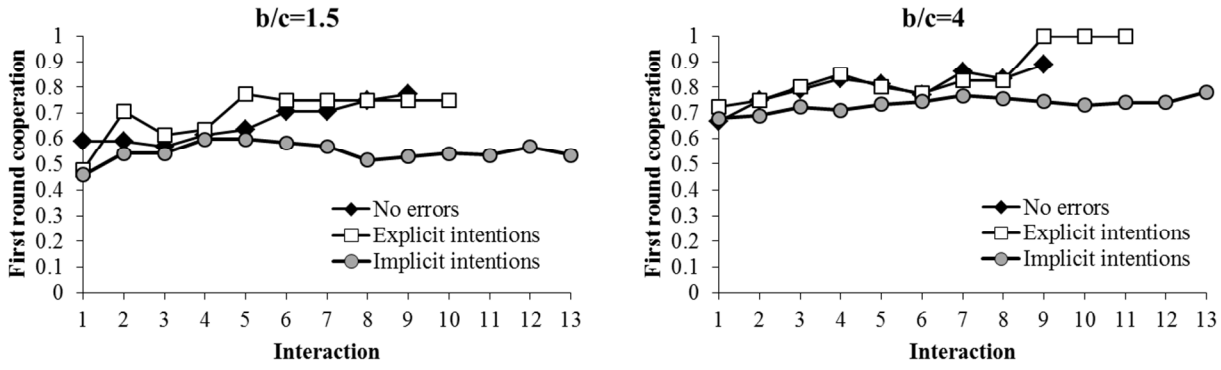


Figure 2. *First period cooperation across supergames for each condition and b/c.*

Before addressing our main questions of interest, we investigate the extent to which behavior changed over the course of a session as the result of learning. Figure 2 displays measures of aggregate behavior in each supergame, and suggests that some learning occurred, especially in the no-error and explicit-intentions conditions. Examining each condition separately, we find a significant increase in first period cooperation over supergame number in the explicit-intentions treatments (b/c=1.5: p=0.010; b/c=4, p<0.001) and the no-error controls (b/c=1.5: p=0.003; b/c=4, p=0.034), but not in the implicit-intentions controls (b/c=1.5: p=0.788; b/c=4, p=0.166).[15] Consistent with this, a regression of all data together[16] shows that learning (as measured by the effect of supergame number on first-period cooperation) is significantly slower in the implicit-intentions control compared to the no-error control (p<0.001), but that learning is equally fast

[15] Logistic regression with robust standard errors clustered on subjects and group (i.e. subject pairing). Two-level clustering in all regressions follows the procedure described in Thompson (2011). See Appendix A Table A1 for full regression table. We note that this clustering accounts for correlation of multiple decisions from the same individual, as well as the correlation in decisions by the two players in a given interaction that arises from their moves later in the interaction being influenced by decisions that occurred earlier in interaction. Following common practice in the experimental literature on repeated games with re-matching, we opted to have relatively few larger sessions in each condition, rather than a greater number of small sessions, in order to minimize contagion effects. Because of this, however, we do not have enough independent sessions to cluster on session. Therefore, our regression analyses are limited by not being able to take into account correlation induced by learning or contagion. Nonetheless, it is encouraging that we see similar results in our SFEM analyses, which do not have this problem.

[16] Logistic regression with robust standard errors clustered on subjects and group, including dummies for condition (no errors, implicit-intentions; explicit-intentions taken as baseline) and b/c ratio, and interacting condition dummies with supergame number. P-values are those associated with the condition dummy X supergame number interaction coefficients. We note that this regression finds significantly more cooperation at b/c=4 than b/c=1.5 (p<0.001), consistent with previous work (Dal Bó 2005, Dreber et al. 2008).

(p=0.615) in the no-error control and the explicit-intentions treatment.[17] To balance the need for data with the evidence of learning, we focus our analysis on the last four supergames of each session, as in Fudenberg et al. (2012).

To address Questions 1 and 2, we now examine the aggregate level of cooperation over the last four supergames, as well as the fraction of the time that subjects cooperated in the first period of a new supergame.

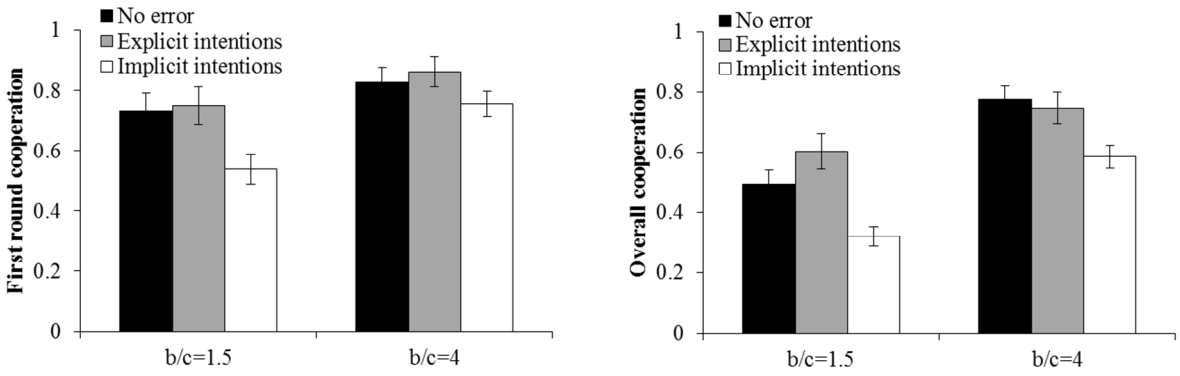


Figure 3. *First period cooperation and overall cooperation for each condition and b/c.*[18]

QUESTION 1 Does observing intentions allow more cooperative play compared to no intentions?

QUESTION 2 How similar are cooperation rates in the explicit-intentions treatments compared to the no-error treatments?

Figure 3 suggests that there is roughly equal cooperation in the explicit-intentions treatments and the no-error controls, and less cooperation in the implicit-intentions controls (Questions 1 and 2): for b/c=1.5, cooperation levels are 60% and 49% in the explicit-intention

[17] These results differ from those of Bereby-Meyer and Roth (2006), who find that learning is slower in a finitely repeated game with probabilistic payoffs where intentions are explicit than in one with deterministic payoffs. Their analysis compares first period cooperation in the first supergame with that of the last one (rather than regressing across all supergames). Analyzing our data in that way does not change our results qualitatively: we still find that learning is significantly faster in explicit-intentions than implicit-intentions (p=0.018) but that there is no significant difference between explicit-intentions and no error (p=0.808). We return to this difference between our results and those of Bereby-Meyer and Roth in the Discussion and Appendix B.

[18] Error bars are generated by linear regression taking action (0=D, 1=C) as the dependent variable, and using robust standard errors clustered on subjects and group. For each set of error bars, we analyze the data from the indicated condition only, and show the standard error associated with the intercept in a regression with no independent variable.

treatment and the no-error control respectively, whereas it is only 32% in the implicit-intentions control. For b/c=4, the corresponding numbers are 75% and 78% versus 59%. Statistical tests confirm that there is indeed more cooperation when intentions are explicit than when they are implicit (significant difference at p=0.02 or less for all comparisons, except for first period cooperation in the cooperative b/c=4 treatment, where the differences with implicit-intentions are not significant, although of the same sign (vs. explicit-intentions, p=0.14; versus no error, p=0.294)).[19]

To answer Questions 3 through 5, we turn from aggregate behavior to considering the particular strategies used by subjects in our experiments. As one way to do this, we use the SFEM of Dal Bó and Frechette (2011) to assign probability weights to a predefined set of strategies. We complement this method with descriptive statistics that do not require the specification of a particular strategy set, but instead make assumptions about the general form of strategies employed.

Before addressing the remaining experimental questions in turn, we briefly describe the SFEM of strategy estimation and present its results. We then draw from these results to answer our questions. We will only summarize this method here (see Dal Bó and Frechette 2011 and Fudenberg et al. 2012 for more information). The idea is to restrict attention to a relatively large but finite set *S* of strategies, and suppose that each subject chooses a fixed element of *S* in the last 4 supergames, and moreover that regardless of whether there are exogenous errors, subjects make mistakes or "mental errors" when choosing their intended action. These mistakes let us assign a positive likelihood to any history for player and any strategy, and we can then assign an aggregate likelihood to any probability distribution *p* on *S*. We estimate *p* by MLE, and compute the standard errors by bootstrap; Appendix C presents the likelihood function we use.

A key aspect of this approach is choosing the set of strategies S to include in the estimation. Given the available data it is not possible to distinguish all possible strategies, as some histories arise only rarely and infinitely many can never occur at all in any finite sample. Guided by theoretical considerations and past empirical work we begin with a set of 38 of

[19] Unless otherwise noted, all subsequent p-values are generated using logistic regressions taking cooperation choice (0=D, 1=C) as the dependent variable and a treatment dummy as the independent variable, with robust standard errors clustered on subjects and group. To generate the p-values reported in this paragraph, we performed pairwise comparisons including the data from each relevant pair of bars in Figure 3. The differences in cooperation are also significant when looking at the last 6 supergames instead of the last 4, and in particular there is the least cooperation in the implicit-intentions treatment, though the levels do change somewhat. See Appendix A Tables A2-A5 for regression details.

strategies, and then discard those that did not seem to be present in at least one payoff specification (including the no-error and implicit-intention controls).[20] Roughly speaking, we start from the strategies that Dal Bó and Frechette (2011) and Fudenberg et al. (2012) found had a non-negligible share in at least one treatment, and add similar ones that condition on intentions alone or both intentions and outcomes. Appendix D lists all of the 38 original strategies and the procedure for discarding strategies.

Our final strategy set includes 17 strategies, which are described in Table 2. In addition to describing each strategy, Table 2 also indicates which strategies are lenient, in that they wait for multiple defections to punish, and which are forgiving, in that they are willing to return to cooperating following a breakdown in cooperation. We are particularly interested in these lenient strategies given their prevalence in the implicit-intentions controls, as reported in Fudenberg et al. (2012).

The probability assigned to these 17 strategies by the SFEM procedure in each of our 6 conditions is shown in Table 3.[21] Only strategies that condition on outcomes are included in the SFEM for the no-error control (because intention and outcome are the same) and in the implicit-intention control (because intention information was unavailable to the subjects).

[20] Because the game has a random termination period we do not include strategies that depend on the number of periods.

[21] We note that similar strategies were estimated to be present in (i) our no-error controls and Dal Bó & Frechette's (2012) no-error games (they use a modified strategy method), and in (ii) our results and the games with no error and with public errors of Aoyagi et al. 2013. This provides evidence of the validity of the SFEM procedure. Note also that in both the observed intentions and no-errors conditions there is a fairly high rate of cooperation throughout every period when b/c=4; this can make it difficult to separate conditionally cooperative strategies from ALLC. For this reason we are not sure how to interpret the difference in the estimated shares of ALLC in these conditions.

Table 2. *Descriptions of the 17 strategies included in the main* SFEM *analysis.*

Strategy	**Abbreviation**	**Description**
Unconditional strategies		
Always Cooperate	ALLC	Always play C (Lenient & forgiving)
Always Defect	ALLD	Always play D
Strategies that condition on outcomes		
Tit-for-Tat	TFT	Play C unless partner's action was D last period (Forgiving)
Tit-for-2-Tats	TF2T	Play C unless partner's action was D in both of the last 2 periods (Lenient & forgiving)
Tit-for-3-Tats	TF3T	Play C unless partner's action was D in all of the last 3 periods (Lenient & forgiving)
2-Tits-for-1-Tat	2TFT	Play C unless partner's action was D in either of the last 2 periods (2 periods of punishment if partner plays D) (Forgiving)
2-Tits-for-2-Tats	2TF2T	Play C unless there were 2 consecutive periods out of the last 3 periods in which either the partner's action was D (2 periods of punishment if partner plays D twice in a row) (Lenient & forgiving)
Grim	Grim	Play C until either player's action is D, then play D forever
Grim 2	Grim2	Play C until 2 consecutive periods occur in which either player's action was D, then play D forever (Lenient)
Grim 3	Grim3	Play C until 3 consecutive periods occur in which either player's action was D, then play D forever (Lenient)
Exploitive Tit-for-Tat	D-TFT	Play D in the first period, then play TFT
Strategies that condition on intentions		
Intention based Tit-for-Tat	TFT-I	Play C unless partner's intention was D last period (Forgiving)
Intention based Tit-for-3-Tats	TF3T-I	Play C unless partner's intention was D in all of the last 3 periods (Lenient & forgiving)
Intention based Grim	Grim-I	Play C until either player's intention is D, then play D forever
Intention based Grim 2	Grim2-I	Play C until 2 consecutive periods occur in which either player's intention was D, then play D forever (Lenient)
Strategies that condition on both intentions and outcomes		
Tolerant Tit-for-Tat	TFT-T	Play C unless partner's intention and action were both D last period (Forgiving)
Punitive 2-Tits-for-2-Tats	2TF2T-P	Play C unless there were 2 consecutive periods out of the last 3 periods in which either the partner's intention or action was D (2 periods of punishment if partner intends or actually plays D twice in a row) (Lenient & forgiving)

Table 3. *SFEM results by condition. Bootstrapped standard errors shown in parentheses. Difference from 0 indicated by **p<0.01, *p<0.05, †p<0.1*

	b/c=1.5			b/c=4		
	No Error	Explicit-Intentions	Implicit-Intentions	No Error	Explicit-Intentions	Implicit-Intentions
ALLC	0	0.02	0	0.24*	0.07	0.06†
	(0)	(0.03)	(0)	(0.1)	(0.06)	(0.03)
TFT	0.27**	0	0.19**	0.14†	0	0.07*
	(0.09)	(0)	(0.05)	(0.07)	(0)	(0.03)
TF2T	0	0.04	0.05	0	0.02	0.2**
	(0.02)	(0.04)	(0.03)	(0.04)	(0.02)	(0.07)
TF3T	0	0	0.01	0	0	0.09†
	(0)	(0.02)	(0.01)	(0.04)	(0.02)	(0.05)
2TFT	0	0	0.06	0.15*	0	0.03
	(0)	(0)	(0.04)	(0.07)	(0)	(0.02)
2TF2T	0.06	0	0	0	0	0.12*
	(0.04)	(0)	(0.02)	(0)	(0.01)	(0.06)
Grim	0.43**	0	0.14**	0.15†	0	0.04
	(0.08)	(0)	(0.04)	(0.08)	(0)	(0.02)
Grim2	0.01	0	0.06†	0.16†	0.04	0.05†
	(0.02)	(0)	(0.03)	(0.08)	(0.04)	(0.03)
Grim3	0	0.01	0.06†	0	0.07	0.11**
	(0)	(0.02)	(0.03)	(0.05)	(0.06)	(0.04)
ALLD	0.18**	0.18**	0.29**	0.07†	0.10*	0.23**
	(0.06)	(0.06)	(0.07)	(0.04)	(0.05)	(0.05)
D-TFT	0.05	0	0.14**	0.09†	0	0
	(0.04)	(0)	(0.05)	(0.05)	(0)	(0)
TFT-I		0.20*			0.04	
		(0.09)			(0.09)	
TF3T-I		0.07			0.16†	
		(0.05)			(0.08)	
Grim-I		0.26**			0.14†	
		(0.08)			(0.08)	
Grim2-I		0.02			0.18†	
		(0.04)			(0.09)	
TFT-T		0.04			0.19†	
		(0.04)			(0.10)	
2TF2T-P		0.15*			0	
		(0.06)			(0.02)	
Gamma	*0.36***	*0.31***	*0.46***	*0.35***	*0.39***	*0.43***
	(0.02)	*(0.02)*	*(0.02)*	*(0.03)*	*(0.04)*	*(0.02)*
Estimated error rate in strategy implementation	6%	4%	10%	5%	7%	9%

We now turn to our remaining experimental questions.

QUESTION 3: How close do subjects come to basing their play solely on intentions?

Examining Table 3, we see that a majority of subjects in the explicit-intentions treatments disregard outcomes (i.e. play unconditional strategies or strategies that condition exclusively on intentions): 77% of probability weight at b/c=1.5, 69% of probability weight at b/c=4. Strategies conditioning exclusively on outcomes account for only 4% at b/c=1.5 and 13% and b/c=4, and strategies that condition on both intentions and outcomes account for 19% in each payoff specification.

As an additional way to examine this question, we consider all histories in which the opponent's intent and actual move in the last period differed. In 82% of such cases (b/c=1.5: 84%, b/c=4: 80%), the subject's decision matched the opponent's intent rather than the opponent's actual move. The results are qualitatively unchanged if we exclude subjects who cooperated in fewer than 25% of all decisions: the subject's decision then matched the opponent's intent in the previous period in 85% of cases (b/c=1.5: 85%, b/c=4: 85%).

Thus both the SFEM and the descriptive statistics suggest that a large majority of subjects base their play solely on intentions. This contrasts with the findings in some experiments on one-shot games, where maximizing money payoffs requires ignoring intentions entirely. Even though many subjects do condition at least partially on intentions in these games, there is much more of a tendency for them to condition on outcome as well compared to our results.[22] The fact that intentions play a larger role in repeated games is consistent with the view that reciprocity in one-shot games stems from the application of a heuristic that was developed for repeated interactions.

We now focus our attention on the subset of players who *do* actually condition on outcomes to answer Question 4.

QUESTION 4: To the extent that subjects condition on realized outcomes as well as intentions in the explicit-intentions treatment, how do they do this?

[22] For example, pooling data from the two treatments of Charness and Levine (2007), 59% of subjects (23 out of 39) rewarded when both intent and outcome were good, while 28.3% (15 out of 53) rewarded when intent was good but outcome was bad. Thus if all the subjects who rewarded after (good intent, bad outcome) would have done so after (good intent, good outcome), 28.3% of subjects used a purely intention-based strategy while 30.7% also conditioned on outcomes. The results of Cushman et al. (2009) are even more extreme, with no subjects conditioning purely on intentions, 36.7% conditioning purely on outcomes, and 46.7% conditioning on both intentions and outcomes.

Considering the results in Table 3, we see that subjects in the two payoff specifications condition on outcomes differently. At b/c=1.5, a non-negligible probability weight (15%) is assigned to the strategy 2TF2T-P. In general, 2TF2T waits for two periods of defection by the partner and then punishes for two periods. The 2TF2T-P variant of 2TF2T is "punitive," in that it uses outcomes to punish an accidental D, but does not use a realization of C to forgive an intended D. That is, these players condition on outcome when the intention was C, but not when the intention was D.[23] At b/c=4, conversely, a non-negligible probability weight (19%) is assigned to the strategy TFT-T (4% of subjects at b/c=1.5 also play TFT-T). This strategy is a variant of TFT that is "tolerant," in that it uses outcomes to forgive would-be defectors who cooperated by accident, but not to punish unintended defections. That is, in contrast to the punitive version of 2TF2T, this tolerant strategy conditions on outcomes when the intention was D, but not when the intention was C.

To complement the SFEM analysis, we use simple descriptive measures that implicitly suppose that subjects ignore observations from two or more periods ago and only condition on observations from the previous period (Figure 4). Consistent with the MLE results, Figure 4 suggests that at b/c=1.5 subjects are punitive and condition on outcome when the opponent's intention was C, but not when the opponent's intention was D. This visual impression is confirmed by a positive relationship between a player's probability of cooperating and the opponent's actual move last period when the opponent intended to play C at b/c=1.5 ($p<0.001$).[24] Again consistent with the SFEM, we see a different pattern at b/c=4; here, subjects are tolerant and condition on outcome when the opponent's intention was D but not C. The relationship between cooperation and opponent's actual move last period when the opponent intended to play D at b/c=4 is not statistically significant ($p=0.236$), although the magnitude of the difference is not insubstantial (36.4%C when outcome is C, 27.7%C when outcome is D).

The tolerant strategy at b/c=4 is somewhat intuitive: here the gains from cooperation are high, so subjects have an incentive to signal a willingness to cooperate when the opponent plays C by accident. The punitive behavior at b/c=1.5 is perhaps less expected, but we can think of two possible explanations. First, an inequity-averse player might punish an accidental defection even at the cost of potentially derailing the cooperative relationship, in order to avoid earning less than

[23] Charness and Levine (2007) find something similar in their one-shot experiments where some subjects only reward if both the intended and actual outcome are good.
[24] See Appendix E Table E1 for regression results.

the opponent. This potential cost increases dramatically with the returns to cooperation, explaining the lower level of punitive play at b/c=4. Second, when b/c=1.5, cooperation is just barely risk dominant, and when a subject sees his opponent accidentally play D, she may worry that the opponent will switch to D thereafter, as the opponent might fear that the accident will be misinterpreted. We don't see how to distinguish these hypotheses with our current data.

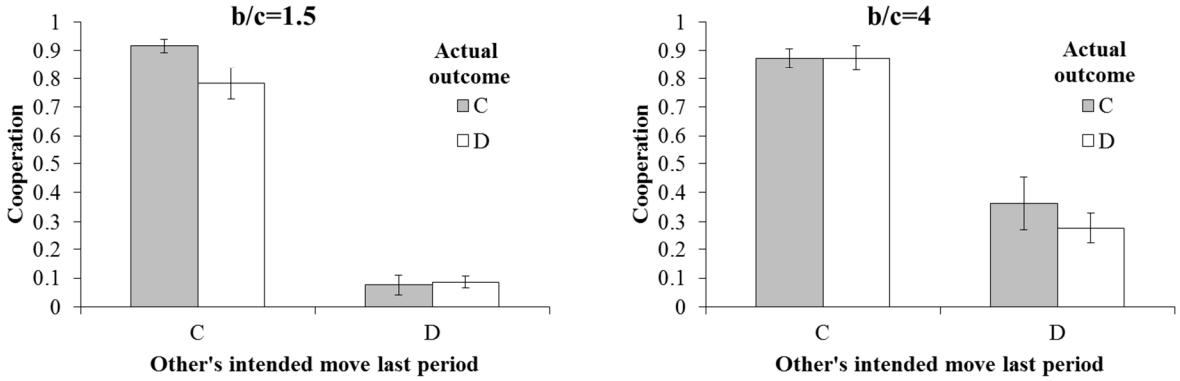


Figure 4. *Probability of cooperating in the last 4 supergames of the explicit-intentions conditions, as a function of opponent's intention and actual move in the previous period.*[25]

A natural question that arises from these results is whether subjects learn to ignore outcomes over time. We do not find evidence of such learning: Interacting opponent's actual move last period with supergame number finds non-significant coefficients trending in the *positive* direction, both for b/c=1.5 when opponent intended C (p=0.094; controlling for period, player's intended action last period, player's overall frequency of cooperation, and player's frequency of cooperation across first periods, p=0.193) and for b/c=4 when opponent intended D (p=0.161; with controls, p=0.218).[26] The conclusion that subjects do not learn to ignore outcome is reinforced by comparing an SFEM analysis on the first four supergames to that on the last four: fewer subjects condition on both intentions and outcomes in early supergames (b/c=1.5: 9%, b/c=4: 4%) than in later supergames (b/c=1.5: 19%; b/c=4: 19%). Thus there is no evidence

[25] Error bars indicate standard errors of the mean clustered on subject and pairing, as in Figure 3. Average behavior over all subjects is shown.

[26] We find qualitatively equivalent results when, instead of assuming that learning is linear in supergame number, we compare play in the first four supergames with play in the last four. (Interacting a "last 4 supergames" dummy with opponent's actual move finds a non-significant positive coefficient in all cases; b/c=1/5, opponent intended C: p=0.13 without controls, p=0.33 with controls; b/c=4, opponent intended D: p=0.174 without controls, p=0.160 with controls).

that reliance on outcomes decreases with experience; if anything, it seems that reliance on outcomes may increase over time. See Appendix E for full regression tables and first four supergame SFEM results.

QUESTION 5: How do the strategies used in the explicit intentions games compare to those used in games with implicit intentions or no errors?

In addition to asking whether and how subjects used realized outcomes in addition to intentions to guide their play, we are interested in how various strategic features of play with explicit intentions compare to those in the no-error and implicit-intentions controls. Specifically, we compare the memory length of the strategies used in the three conditions, and the extent to which play is "lenient" in the sense of not punishing the first deviation by an opponent. Intuitively, the high share of lenient strategies observed in the implicit-intentions controls may correspond to subjects giving their partner the benefit of the doubt that a defection could have occurred by accident, and to combine this sort of leniency with punishment for persistent defections requires strategies to look back more than one period.[27] Thus we might expect less leniency and more simple strategies in the explicit-intentions conditions compared to the implicit-intentions conditions.

Table 4 shows the relevant aggregated SFEM frequencies, as well as various descriptive statistics. First we consider the SFEM aggregations. In terms of strategy complexity, we see that the explicit-intentions treatments look very similar to the no-error controls: a large majority of cooperative strategies are simple in that they are either unconditional or condition/trigger based on the previous period only (b/c=1.5: 74% explicit-intentions, 75% no error; b/c=4, 81% explicit-intentions, 77% no error). This stands in stark contrast to the implicit-intentions controls, where the frequency of simple strategies is cut nearly in half (43% at both b/c=1.5 and b/c=4).

Next we consider leniency. We know from Fudenberg et al. (2012) that when intentions are implicit, leniency is common (and very successful) at b/c ratios where cooperative equilibria exist (such as b/c=4), but relatively rare at the low b/c ratio of 1.5 where there are no cooperative equilibria. If leniency reflects an attempt to infer the intentions of one's partner, we would expect

[27] Fudenberg et al. (2012) also considers the strategic element of 'forgiveness,' or willingness to return to cooperate after punishing a defection. Unlike leniency, there is not a clear a priori prediction about the effect of observing intentions on forgiveness. Thus we do not analyze forgiveness here, but include it in the Appendix F for completeness, where we show that there is not a clear relationship between it and whether intentions are observable.

less leniency at b/c=4 when intentions are explicit than when they are implicit. Consistent with that expectation, at b/c=4, the fraction of cooperative strategies that are lenient in the explicit-intentions condition is much more similar to the no-error control than the implicit-intentions control. When b/c=1.5, it is not clear what to expect, because revealing intentions creates cooperative equilibria where none existed with implicit intentions: while there is less need to infer intentions when intentions are explicit, it is also much less costly to be lenient (since most others are cooperative). We find that when b/c=1.5, leniency in the explicit-intentions treatment is similar to the implicit-intentions control (and actually slightly higher), and both are higher than the no-error control.[28]

We now complement these results with descriptive statistics. For maximum comparability, these measures use intentions in the explicit-intentions treatments, and outcomes in the implicit-intentions and no-error controls. In each case, we measure leniency by examining all histories in which C (either intentional or realized, depending on the measure) occurred in all but the previous period, while in the previous period one subject played D.[29] We then ask how frequently the subject who had hitherto cooperated showed leniency by continuing to cooperate. The results are similar to the SFEM. At b/c=4, leniency in the explicit-intentions condition is lower than the implicit-intentions control, whereas at b/c=1.5, the amount of leniency is similar in explicit-intentions and implicit-intentions.

In sum, we find evidence that in the presence of errors, making intentions explicit increases the frequency of cooperative strategies, reduces the complexity of those cooperative strategies, and also reduces the extent of leniency (at least at b/c=4, the specification in which leniency is common when intentions are implicit).

[28] As the fraction of cooperative strategies varies across condition (in particular, at b/c=1.5 the implicit-intention condition is much lower than the other conditions), we report the fraction of cooperative strategies that are lenient, rather than the fraction of all strategies that are lenient. For completeness, we report the un-normalized values here: b/c-1.5: no-error=7%, explicit-intentions=31%, implicit-intentions=17%; b/c=4: no-error=40%, explicit-intentions=53%, implicit-intentions=63%.

[29] We also include second round decisions in which the first round's outcome was CD.

Table 4. *Aggregated SFEM frequencies and descriptive results by condition.*

	b/c=1.5			b/c=4		
	No error	Explicit intentions	Implicit intentions	No error	Explicit intentions	Implicit intentions
Cooperative strategies in SFEM	77%	82%	56%	84%	90%	77%
% of Cooperative strategies in SFEM that are:						
Memory at most 1	75%	74%	43%	77%	81%	43%
Lenient	9%	38%	31%	47%	59%	81%
Descriptive statistics						
%C first period	73%	75%	54%	83%	86%	76%
%C all periods	49%	60%	32%	78%	75%	59%
Leniency	15%	28%	29%	42%	55%	66%

4. Discussion

We begin by asking how well subjects did in terms of maximizing their payoffs, both overall and by type of strategy used. This provides some insight into which sorts of strategies were (ex post) mistakes, and gives us a rough sense of how close the distribution of play is to an equilibrium - e.g. what percentage of players are receiving close to the best possible payoff given the distribution of play.

Table 5 shows the expected payoff of each strategy given the distribution estimated by the SFEM.[30] In the explicit-intentions treatment at b/c=1.5, the two most prevalent strategies are TFT-I and Grim2-I; these purely intention-based strategies also yield the highest payoff of 5.5. The payoff of the punitive 2TF2T-P, which was also somewhat common, was slightly lower, but this difference is not statistically significant.

[30] This analysis assumes that the SFEM accurately identified the frequency of each strategy. To partially address potential errors in the SFEM identification, we bootstrap standard errors for each expected payoff by repeatedly sampling with replacement from player histories, and then recalculating the SFEM and resulting expected payoff for each strategy.

Table 5. *SFEM frequency and expected payoffs for each strategy in each condition. Highest payoff strategies, and strategies with payoffs that are not statistically different from the highest payoff (based on bootstrapped standard errors using a significance level of p<0.1), are highlighted in gray.*

	b/c=1.5						b/c=4					
	No Error		Explicit Intentions		Implicit intentions		No error		Explicit Intentions		Implicit intentions	
	MLE	Payoff	MLE	Payoff	MLE	Payoff	MLE	Payoff	MLE	Payoff	MLE	Payoff
ALLC		3.5	0.02	3.6		-1.3	0.24	43.0	0.07	36.5	0.06	28.1
TFT	0.27	5.9		4.6	0.19	2.4	0.14	42.3		33.6	0.07	29.0
TF2T		5.7	0.04	4.8	0.05	1.5		43.9	0.02	36.7	0.20	29.6
TF3T		5.5		4.6	0.01	0.9		43.8		37.0	0.09	29.5
2TFT		5.8		4.7	0.06	2.9	0.15	40.8		30.3	0.03	27.0
2TF2T	0.06	5.7		4.9		1.9		43.9		36.3	0.12	29.6
Grim	0.43	5.8		4.4	0.14	3.0	0.15	40.8		25.9	0.04	24.0
Grim2	0.01	5.7		4.8	0.06	2.4	0.16	43.9	0.04	33.5	0.05	27.9
Grim-3		5.5	0.01	4.9	0.06	1.8		43.8	0.07	36.4	0.11	29.2
ALLD	0.18	2.5	0.18	4.1	0.29	3.7	0.07	21.1	0.10	19.1	0.23	21.0
D-TFT	0.05	2.7		4.1	0.15	2.9	0.09	25.4		29.8		28.7
TFT-I			0.20	5.5					0.04	37.7		
TF3T-I			0.07	5.1					0.16	37.4		
Grim-I			0.26	5.5					0.14	37.6		
Grim2-I			0.02	5.3					0.18	37.5		
TFT-T			0.04	5.3					0.19	37.5		
2TF2T-P			0.15	5.2						36.5		

The most common poorly performing strategy here is ALLD. The payoff loss to ALLD is even higher in the no-error control, due to the smaller share of lenient strategies and the increased share of the unforgiving strategy set Grim and its variants. Conversely, ALLD yields the highest payoff in the implicit-intentions control, where it is also the most commonly used strategy. Mistaken extrapolation from that case could help explain the play of ALLD in the explicit-intentions and no-error treatments.

At b/c=4, in the explicit-intentions treatment most subjects play some sort of conditional cooperation strategy based on intentions only or intentions and outcomes; all of these strategies do fairly well, earning payoffs not statistically different from the highest performing strategy. In particular the expected payoff of the tolerant strategy TFT-T is statistically indistinguishable from that of TFT-I. Once again, the most common "mistake" is to play ALLD, which yields a payoff of about 19 versus the high-30's payoffs obtained with conditional cooperation. Thus in

both payoff treatments, a high fraction of the subjects do quite well, and subjects who condition on outcomes as well as intentions do so at essentially no cost.

As shown above in Figure 2, learning is significantly slower in the implicit-intentions control compared to the no-error control, but that learning is equally fast in the no-error control and the explicit-intentions treatment. This latter fact contrasts with the finding of Bereby-Meyer and Roth (2006), who found that adding a stochastic shock to payoffs (as opposed to actions) resulted in slower learning. One possible explanation is that our procedure speeds learning because it focuses attention on the opponent's intentions, which are what subjects need to be learning about in order to reach the best equilibrium.[31] This focus on intentions might be due either to the fact that implementation errors alter both players' payoffs in our case while the lotteries in Bereby-Meyer and Roth were independent, or because the framing of our game suggests to subjects that intentions are what matters.

To directly test this latter possibility, we ran a follow-up experiment on Amazon Mechanical Turk (Horton et al. 2011), recruiting 96 subjects and randomizing them into one of two ways to explain the structure of the random errors, the "Error" and "Lottery" conditions, where we asked subjects about the intentionality of a D outcome when C was chosen. In the Error condition, the probabilistic mechanism was explained with the same language as in our explicit-intentions condition. In the Lottery condition, the probabilistic mechanism was instead explained by saying that there were two options as in the Bereby-Meyer and Roth study.[32]

As predicted, subjects in the Error condition thought that the D outcome was significantly less intentional than subjects in the Lottery condition did (mean intentionality ratings on 1-7 scale: Error: 2.27, Lottery: 3.33; Rank-sum, p=0.036; Tobit regression with robust standard errors: p=0.020; including controls for age, gender and education: p=0.009). This result supports our hypothesis that framing noise as execution errors emphasizes the 'accidental' nature of bad outcomes relative to framing noise as a lottery, and so increases the subjects' attention to the intentions of their partner. Put differently, the execution-error framing may decrease subjects'

[31] Another possibility is that the difference in expected payoffs between cooperative and non-cooperative strategies may have been smaller in Bereby-Meyer and Roth's experiments than in our explicit-intentions conditions, thus providing a weaker signal for learning. To evaluate this possibility, one could use SFEM to estimate the distribution of strategies in Bereby-Meyer and Roth's data and then calculate the expected payoffs of each strategy. However, given that they used fixed length games, it is not clear to us which strategies should be included in the SFEM (e.g. strategies which open with cooperation but then switch to defection after some number of periods) so we do not explore this possibility here.

[32] See the Online Appendix for the full instructions.

sense of 'causal control' (Cushman et al. 2009) relative to the lottery framing, in that the error frame makes it seem as though some other agent (the computer) is causing the bad outcome, rather than the player.[33] We conjecture that this increased the subjects' ability to focus on the intended good outcome and that this is why learning proceeded more quickly.

We also see that subjects use simpler, lower-memory strategies with explicit intentions than when intentions are implicit. This suggests that the more complex strategies found in the implicit-intentions conditions (reported in Fudenberg et al. 2012) use longer memory in part as a way to attempt to learn and track the intentions of other subjects. This is particularly true for b/c=4, where there is a high return to cooperation, and long memory lenient cooperative strategies were most prevalent with implicit intentions.

To further investigate strategic complexity, we examine how response times vary with play and condition.[34] We see that faster decisions are more cooperative ($p<0.001$).[35] Considering variation by condition, we would predict based on the complexity of the decision setting that decision times should be fastest in the no-error control, slowest in the implicit-intentions control, and intermediate with explicit intentions. When we examine the data, we see that the decision times conform to this prediction when b/c=4.[36] However, when b/c=1.5, decision time is longest in the explicit-intentions treatment, with the no-error treatment coming second.[37]

We also see interesting differences across conditions in learning: In the no-error and explicit-intentions conditions, decision times decrease with experience as measured by supergame number on ($p<0.001$ for both), as we would expect; but with implicit intentions,

[33] See also Bolton et al. (2005) who explore procedural fairness versus outcome fairness.

[34] The only paper we are aware of that considers the correlation between response times and cooperation in repeated PDs with noise is Rand et al. (2012), who re-analyze the data of Fudenberg et al. (2012) and find a positive correlation.

[35] Logistic regression with robust standard errors clustered on subject and pairing, considering the last four supergames. Log-10 transformed response time is taken as the independent variable, and controls for condition (dummies for explicit-intention, no-error or implicit-intention), b/c, supergame number and period number are included. Decision times are log-10 transformed as in Rand et al. (2012) to account for the heavily skewed nature of the response time distributions. Equivalent results are found when using untransformed response times, when including all supergames, or both. See Online Appendix A Table OA1.

[36] Log-10 transformed response times: no-error, 0.144; explicit-intentions, 0.185; implicit-intentions, 0.263. An equivalent ordering is obtained when controlling for supergame, period, and whether the decision was C or D.

[37] Log-10 transformed response times: no-error, 0.294; explicit-intentions, 0.224; implicit-intentions, 0.235. An equivalent ordering is obtained when controlling for supergame period, and whether the decision was C or D, except that the very similar response times of explicit-intentions and implicit-intentions flip.

decision times actually *increase* ($p<0.001$).[38] Further investigation of this increasing response time in the implicit-intentions condition finds a significant negative interaction between supergame number and subjects' overall frequency of cooperation when predicting response time ($p<0.001$; controlling for the period 1 intention of the partner in the previous supergame and current partner's action in the previous period, $p=0.001$)[39]: the less often a subject cooperated overall, the more her response time tended to increase with experience (controlling for whether the current decision was C or D). As a result, regressing decision time against supergame number finds no significant relationship when only examining subjects who were largely cooperative (i.e. cooperated in more than 2/3 of all decisions, $p=0.890$).[40] This increase in reaction times by non-cooperative subjects may reflect the effects of learning: these subjects are generally inclined to defect, but over time gather evidence that cooperation might be in their interest. This makes cooperation somewhat more attractive, moving the expected utility of cooperation and defection closer together, and as a result faces these subjects with a more difficult and time consuming choice (in contrast to the initially cooperative subjects, whose inclination is reinforced by experience). Such decision conflict can increase decision times (Evans et al. 2015).

Although we do not have a theoretical explanation for all of the response time data, we believe that the connection between response time and choice of strategy, and how this varies with the strategic environment, is an interesting topic that merits future study.

Conclusion

We conclude that making intentions explicit allows subjects to achieve the same level of cooperation under errors as if errors were not present, because subjects largely ignore outcomes

[38] Linear regressions with robust standard errors clustered on subject and pairing, taking log10(response time) as the DV, supergame as the IV, and including controls for b/c, period and whether the decision was C or D. See Online Appendix A Table OA2..

[39] Linear regression with robust standard errors clustered on subject and pairing, considering the implicit-intentions condition, taking log10(response time) as the DV, supergame number and frequency of cooperation over all decisions as the IVs, and including an interaction between these two terms as well as controls for b/c, period and whether the decision was C or D.

[40] The cutoff of 2/3 was determined by testing at which frequency of overall cooperation the net coefficient on supergame number became non-significant. P-value reflects linear regression with robust standard errors clustered on subject and pairing, considering the implicit-intentions condition, taking log10(response time) as the DV, supergame number and frequency of cooperation over all decisions as the IVs, and controls for b/c, period and whether the decision was C or D. The finding of no effect among these subjects persists when also controlling for the period 1 intention of the partner in the previous supergame and current partner's action in the previous period ($p=0.192$).

and condition only on intentions. This finding is consistent with the predictions of theory in the sense that the highest payoff equilibrium ignores outcomes. Moreover, intention-based strategies are both common and earn high payoffs given the observed distribution of play. Thus institutions that increase the observability of intentions may help to mitigate the negative consequences of errors: when your aim is true, accidents will by and large be forgiven and forgotten.

Acknowledgments

National Science Foundation Grant SES-0954162, the John Templeton Foundation, the Jan Wallander and Tom Hedelius Foundation (Svenska Handelsbankens Forskningsstiftelser) and the Knut and Alice Wallenberg Foundation provided financial support. We are grateful for comments from Yoella Bereby-Meyer, Fiery Cushman, Guillaume Frechette, Moshe Hoffman, Alexander Peysakhovich, Bjørn-Atle Reme, Madison Storm, and Sevgi Yuksel, and we thank Yoella Bereby-Meyer and Al Roth for sharing their data with us.

References

Abreu, Dilip, David Pearce, and Ennio Stacchetti. 1990. "Toward a Theory of Discounted Repeated Games with Imperfect Monitoring." *Econometrica* 58 (5): 1041–63.

Ambrus, Attila, and Ben Greiner. 2012. "Imperfect Public Monitoring with Costly Punishment - An Experimental Study." *American Economic Review* 102 (7): 3317-32

Andreoni, James, Paul M. Brown, and Lise Vesterlund. 2002. "What Makes an Allocation Fair? Some Experimental Evidence." *Games and Economic Behavior* 40, 1–24.

Aoyagi, Masaki, V. Bhaskar, and Guillaume Frechette (2013), "The Impact of Monitoring in Infinitely Repeated Games: Perfect, Public, and Private." Manuscript in preparation.

Aoyagi, Masaki, and Guillaume Frechette. 2009. "Collusion as Public Monitoring Becomes Noisy: Experimental Evidence." *Journal of Economic Theory* 144 (3): 1135–65.

Bereby-Meyer, Yoella, and Alvin E. Roth. 2006. "The Speed of Learning in Noisy Games: Partial Reinforcement and the Sustainability of Cooperation." *American Economic Review* 96(4) 1029-1042.

Bigoni, Maria, Jan Potters, and Giancarlo Spagnolo. 2012. "Flexibility and Collusion with Imperfect Monitoring." Working Paper.

Blonski, Matthias, Peter Ockenfels, and Giancarlo Spagnolo. 2011. "Equilibrium Selection in the Repeated Prisoner's Dilemma: Axiomatic Approach and Experimental Evidence." *American Economic Journal: Microeconomics* 3 (3): 164–92.

Blount, Sally. 1995. “When Social Outcomes aren't Fair: The Effect of Causal Attributions on Preferences.” *Organizational Behavior and Human Decision Processes* 63: 131–144.

Bolton, Gary E., Jordi Brandts, and Axel Ockenfels. 2005. “Fair Procedures: Evidence from Games Involving Lotteries.” *Economic Journal* 115: 1054-1076.

Brandts, Jordi, and Carles Solà. 2001. “Reference Points and Negative Reciprocity in Simple Sequential Games.” *Games and Economic Behavior* 36: 138–157.

Charness, Gary, and David I. Levine. 2007. “Intention and Stochastic Outcomes: An Experimental Study.” *Economic Journal* 117: 1051-1072.

Cushman, Fiery, Anna Dreber, Ying Wang, and Jay Costa. 2009. “Accidental Outcomes Guide Punishment in a ’Trembling Hand’ Game.” *PLoS ONE* 8, 1–7.

Dal Bó, Pedro. 2005. “Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games.” *American Economic Review* 95 (5): 1591–1604.

Dal Bó, Pedro, and Guillaume R. Frechette. 2011. “The Evolution of Cooperation in Infinitely Repeated Games: Experimental Evidence.” *American Economic Review* 101 (1): 411–29.

Dal Bó, Pedro, and Guillaume R. Frechette. 2012. “Strategy Choice In The Infinitely Repeated Prisoners Dilemma.” Working paper.

Dal Bó, Pedro, and Guillaume R. Frechette. 2015. "On the Determinants of Cooperation in Infinitely Repeated Games: A Survey." Forthcoming in the *Journal of Economic Literature.*

Dreber, Anna, David G. Rand, Drew Fudenberg, and Martin A. Nowak. 2008. “Winners Don’t Punish.” *Nature* 452: 348–351.

Evans, Anthony M., Kyle D. Dillon, David G. Rand. 2015 “Decision Conflict and Reflection in Social Dilemmas: Extreme Responses are Fast, But Not Intuitive.” Available at SSRN: http://ssrn.com/abstract=2436750.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2003. ”On the Nature of Fair Behavior.” *Economic Inquiry* 41, 20–26.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2008. “Testing Theories of Fairness — Intentions Matter.” *Games and Economic Behavior* 62: 287–303.

Fischbacher, Urs. 2007. “Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments.” Experimental Economics 10 (2): 171–78.

Fudenberg, Drew, David G. Rand, and Anna Dreber 2012. “Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World.” *American Economic Review* 102 (2), 720-749.

Fudenberg, Drew, David K. Levine, and Eric Maskin. 1994. “The Folk Theorem with Imperfect Public Information.” *Econometrica* 62 (5): 997–1039.

Green, Edward J., and Robert H. Porter. 1984. “Noncooperative Collusion under Imperfect Price Information.” *Econometrica* 52 (1): 87–100.

Kunreuther, Howard, Gabriel Silvasi, Eric Bradlow, and Dylan Small. 2009. “Bayesian Analysis of Deterministic and Stochastic Prisoner's Dilemma Games.” *Judgment and Decision Making* 4 (5): 363–384.

Horton John J., David G. Rand, and Richard J Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14 (3): 399–425.

McCabe, Kevin A., Mary L. Rigdon, and Vernon L. Smith. 2003. "Positive Reciprocity and Intentions in Trust Games." *Journal of Economic Behavior and Organization* 52: 267–275.

Radner, Roy, Roger Myerson, and Eric Maskin. 1986. "An Example of a Repeated Partnership Game with Discounting and with Uniformly Inefficient Equilibria." *Review of Economic Studies* 53: 59-69.

Rand, David G., Joshua D. Greene, and Martin A. Nowak. 2012. "Spontaneous Giving and Calculated Greed." *Nature* 489: 427–430.

Rand, David G., and Martin A. Nowak. 2013. "Human cooperation." *Trends in Cognitive Sciences* 17: 413-425.

Rubin, Jared, and Roman Sheremeta. Forthcoming. "Principal-Agent Settings with Random Shocks. *Management Science*.

Schächtele, Simeon, Tobias Gerstenberg, and David Lagnado. 2011. "Beyond Outcomes: The Influence of Intentions and Deception." Working Paper, UCL.

Thompson, Samuel B. 2011. "Simple formulas for standard errors that cluster by both firm and time." *Journal of Financial Economics* 99: 1-10.

Van Lange, Paul A. M., Jaap W. Ouwerkerk, and Mirjam J. A. Tazelaar. 2002. "How to overcome the detrimental effects of noise in social interaction: The benefits of generosity." *Journal of Personality and Social Psychology* 82: 768-780.

Appendix A:

Table A1. First period cooperation by supergame. Logistic regression with robust standard errors clustered on subject and pairing.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Explicit-intentions		No-error		Implicit-intentions		All data
	b/c=1.5	b/c=4	b/c=1.5	b/c=4	b/c=1.5	b/c=4	
Supergame	0.118***	0.189***	0.118***	0.129**	0.00583	0.0336	0.123***
	(0.0456)	(0.0262)	(0.0396)	(0.0607)	(0.0217)	(0.0243)	(0.0332)
Explicit							0.0307
							(0.305)
Implicit							0.0260
							(0.263)
Explicit X Supergame							0.0230
							(0.0458)
Implicit X Supergame							-0.101***
							(0.0369)
b/c							0.318***
							(0.0834)
Constant	0.214	0.658*	0.0841	0.780***	0.152	0.793***	-0.439
	(0.269)	(0.344)	(0.298)	(0.300)	(0.218)	(0.226)	(0.310)
Observations	416	392	396	396	810	1,072	3,482
Subject-clusters	44	40	44	48	72	90	338
Pairing-clusters	208	196	198	198	405	536	1,741

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A2. Mean comparisons, overall C, comparing no-error with explicit intentions (columns 1-2), comparing implicit intentions with explicit intentions (columns 3-4), comparing implicit intentions with no-error (columns 5-6), for last 4 interactions. Logistic regression with robust standard errors clustered on subject and pairing.

	(1)	(2)	(3)	(4)	(5)	(6)
	Comparison: Explicit		Comparison: Explicit		Comparison: No error	
	b/c=1.5	b/c=4	b/c=1.5	b/c=4	b/c=1.5	b/c=4
No error	-0.444	0.158				
	(0.316)	(0.375)				
Implicit			-1.167***	-0.735**	-0.722***	-0.893***
			(0.288)	(0.317)	(0.245)	(0.303)
Constant	0.415*	1.083***	0.415*	1.083***	-0.0294	1.241***
	(0.247)	(0.273)	(0.247)	(0.273)	(0.196)	(0.257)
Observations	2,972	3,000	4,032	4,444	4,052	4,708

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A3. Mean comparisons, first round C, comparing no-error with explicit intentions (columns 1-2), comparing implicit intentions with explicit intentions (columns 3-4), comparing implicit intentions with no-error (columns 5-6), for last 4 interactions. Logistic regression with robust standard errors clustered on subject and pairing.

	(1)	(2)	(3)	(4)	(5)	(6)
	Comparison: Explicit		Comparison: Explicit		Comparison: No error	
	b/c=1.5	b/c=4	b/c=1.5	b/c=4	b/c=1.5	b/c=4
No error	-0.0889	-0.264				
	(0.441)	(0.550)				
Implicit			-0.946**	-0.708	-0.857**	-0.444
			(0.392)	(0.480)	(0.356)	(0.423)
Constant	1.099***	1.836***	1.099***	1.836***	1.010***	1.572***
	(0.333)	(0.420)	(0.332)	(0.420)	(0.289)	(0.354)
Observations	352	352	464	520	464	552

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A4. Mean comparisons, overall C, comparing no-error with explicit intentions (columns 1-2), comparing implicit intentions with explicit intentions (columns 3-4), comparing implicit intentions with no-error (columns 5-6), for last 6 interactions. Logistic regression with robust standard errors clustered on subject and pairing.

	(1)	(2)	(3)	(4)	(5)	(6)
	Comparison: Explicit		Comparison: Explicit		Comparison: No error	
	b/c=1.5	b/c=4	b/c=1.5	b/c=4	b/c=1.5	b/c=4
No error	-0.414	0.132				
	(0.294)	(0.338)				
Implicit			-1.202***	-0.712**	-0.788***	-0.844***
			(0.264)	(0.285)	(0.231)	(0.274)
Constant	0.462**	1.024***	0.462**	1.024***	0.0482	1.156***
	(0.227)	(0.246)	(0.227)	(0.245)	(0.187)	(0.232)
Observations	4,332	4,328	6,046	6,512	6,026	6,888

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A5. Mean comparisons, first round C, comparing no-error with explicit intentions (columns 1-2), comparing implicit intentions with explicit intentions (columns 3-4), comparing implicit intentions with no-error (columns 5-6), for last 6 interactions. Logistic regression with robust standard errors clustered on subject and pairing.

	(1)	(2)	(3)	(4)	(5)	(6)
	Comparison: Explicit		Comparison: Explicit		Comparison: No error	
	b/c=1.5	b/c=4	b/c=1.5	b/c=4	b/c=1.5	b/c=4
No error	-0.206	-0.324				
	(0.419)	(0.499)				
Implicit			-0.872**	-0.649	-0.666*	-0.325
			(0.377)	(0.450)	(0.345)	(0.386)
Constant	1.039***	1.768***	1.039***	1.768***	0.833***	1.444***
	(0.315)	(0.388)	(0.315)	(0.388)	(0.276)	(0.312)
Observations	528	528	696	780	696	828

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Appendix B: Re-analysis of learning in Bereby-Meyer and Roth (2006)

Bereby-Meyer and Roth (2006) (BMR) compare learning rates in finitely repeated games where payoffs are either deterministic or probabilistic.[41] They have also include a 'sun-spots' control in which payoffs are deterministic, but players are also presented with the outcome of two random lotteries (that do not effect payoffs) each turn. They conclude that learning to cooperate in period 1 occurs more slowly in the probabilistic condition than in either the deterministic or sun-spot conditions. As the main text explains, they used a different analysis strategy. Instead of regressing first period cooperation against supergame number, BMR compared just the first and the last supergame. This difference in methods did not change the analysis of our data.

Now, we re-analyze their data using our learning metric: comparing the coefficient for supergame number when predicting first period cooperation across conditions.[42] Doing so, we also find that there is significantly slower learning when payoffs are probabilistic compared to deterministic in a repeated game (condition X supergame: coeff = 0.120, $p<0.001$), but we do not find a significant difference in learning speed between their probabilistic condition and their sun-spot control (coeff = 0.032, $p=0.224$). As can be seen in Figure B1, cooperation in the sun-spot control climbs rapidly in supergames 2 and 3, but then stabilizes.

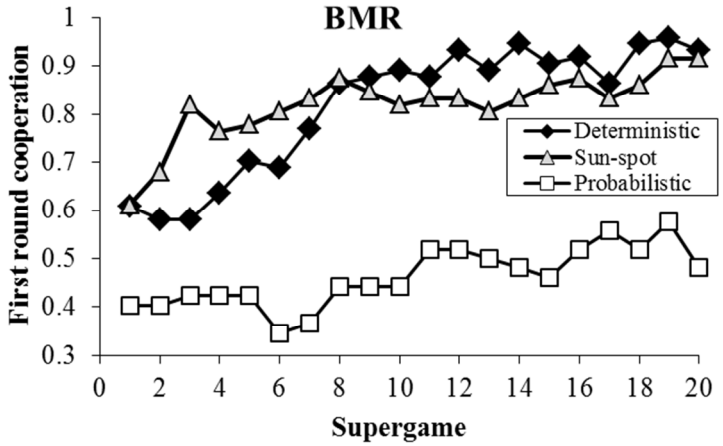


Figure B1. Cooperation in period 1 of the finitely repeated games from Bereby-Meyer and Roth (2006).

[41] They also study one-shot games.

[42] As in the main text, we use logistic regression with standard errors clustered on subject and group.

If we restrict our re-analysis to supergames 1 and 20 (as in their original analysis), we replicate their result: significantly more change from supergame 1 to 20 in both the deterministic ($p=0.006$) and the sun-spot ($p=0.005$) conditions compared to the probabilistic condition. These two methods of analysis give different results in the sun-spots condition because most of the learning there occurs in the first three supergames; regressing over all supergames shows relatively little change in first period cooperation per supergame, despite that fact that first period cooperation increases substantially from the first supergame to the last.

Overall, it seems that the noise in BMR had a substantially different effect on learning than the shocks to actions in our experiments. One possible explanation is that the noise in their probabilistic condition psychologically feels different from how we introduce noise.[43]. Although these two processes are mathematically equivalent, our procedure places more emphasis on the role of intentions: it makes it feel like the other person didn't *mean* to choose the outcome in cases were errors occur. In BMR's setup, although the players do not have direct control over the outcomes, it may not feel to the opponent that the randomness of lottery is changing the actor's intent. As discussed in the main text, we explore this possibility with an additional experiment.

[43] Recall that in their probabilistic condition, players select between two lotteries, and then an outcome is drawn from the chosen lottery. In our explicit-intention conditions, players choose a fixed outcome, and the computer then switches their choice to the opposite with some probability.

Appendix C: Details of the Structural Frequency Estimation Method (SFEM)

We use the SFEM introduced by Dal Bo & Frechette (2011) and used in Fudenberg et al. (2012). We suppose that if subject *i* uses strategy *s,* her chosen action in round *r* of supergame *k* is C if $s_{ikr}(s)+\gamma\varepsilon_{ikr} \geq 0$, where $s_{ikr}(s)=1$ if strategy *s* says to play C in round *r* of supergame *k* given the history to that point, and $s_{ikr}(s)=-1$ if *s* says to play D. Here ε_{ikr} is an error term that is independent across subjects, rounds, supergames, and histories, γ parameterizes the probability of mistakes, and the density of the error term is such that the overall likelihood that subject *i* uses strategy *s* is

$$(1) \qquad p_i(s) = \Pi_k \Pi_r \left(\frac{1}{1+\exp(-s_{ikr}(s)/\gamma)}\right)^{y_{ikr}} \left(\frac{1}{1+\exp(s_{ikr}(s)/\gamma)}\right)^{1-y_{ikr}},$$

where y_{ikr} is 1 if the subject chose C and 0 if the subject chose D. [44]

For any given set of strategies S and proportions p¸ we then derive the likelihood for the entire sample as a mixture model, namely $\sum_I \ln\left(\sum_{s\in S} p(s)p_i(s)\right)$. Note that the specification assumes that all subjects are ex-ante identical with the same probability distribution over strategies and the same distribution over errors; one could relax this at the cost of adding more parameters. Because *p* describes a distribution over strategies, this likelihood function implies that in a very large sample we expect fraction *p(s)* of subjects to use strategy s, though for finite samples there will be a non-zero variance in the population shares. We use maximum likelihood estimation (MLE) to estimate the prevalence of the various strategies, and bootstrapping to associate standard errors with each of our frequency estimates. We construct 100 bootstrap samples for each treatment by randomly sampling the appropriate number of subjects with replacement. We then determine the standard deviation of the MLE estimates for each strategy frequency across the 100 bootstrap samples. The validity of this procedure was demonstrated using simulated data in Fudenberg et al 2012.

[44] Thus the probability of an error in implementing one's strategy is 1/(1+exp(1/γ)). Note that this represents error in intention, rather than the experimentally imposed error in execution. This formulation assumes that all strategies have an equal rate of implementation error. In Fudenberg et al. (2012) we show that the MLE estimates of strategy shares are robust to allowing each strategy have a different value of γ.

Appendix D: SFEM including all intent-outcome hybrids (only the explicit-intentions conditions)

	b/c=1.5	b/c=4
ALLC	0.02	0.07
	(0.03)	(0.06)
TFT	0	0
	(0)	(0)
TF2T	0	0
	(0.01)	(0)
TF3T	0	0
	(0)	(0)
2TFT	0	0
	(0)	(0)
2TF2T	0	0
	(0)	(0)
G	0	0
	(0)	(0)
G2	0	0.04
	(0)	(0.04)
G3	0	0.07
	(0)	(0.06)
ALLD	0.18**	0.1*
	(0.06)	(0.05)
DTFT	0	0
	(0)	(0)
TFTI	0.19*	0.03
	(0.09)	(0.09)
TF2TI	0.02	0
	(0.04)	(0.02)
TF3TI	0	0.16†
	(0.03)	(0.09)
2TFTI	0.09	0
	(0.07)	(0.01)
2TF2TI	0	0
	(0.04)	(0)
GI	0.24**	0.15†
	(0.08)	(0.09)
G2I	0	0.17
	(0.01)	(0.11)
G3I	0	0
	(0.01)	(0.01)
D-TFTI	0.02	0
	(0.03)	(0)

	b/c=1.5	b/c=4
TFTI CC	0.01	0
	(0.02)	(0)
TF2TI CC	0	0.03
	(0.03)	(0.04)
TF3TI CC	0.08	0
	(0.06)	(0.02)
2TFTI CC	0	0
	(0)	(0)
2TF2TI CC	0.08	0
	(0.06)	(0.01)
GI CC	0	0
	(0)	(0)
G2I CC	0	0
	(0)	(0)
G3I CC	0	0
	(0)	(0.04)
D-TFTI CC	0	0
	(0)	(0)
TFTI DD	0	0.18†
	(0.04)	(0.1)
TF2TI DD	0.06	0
	(0.05)	(0.01)
TF3TI DD	0	0
	(0)	(0.05)
2TFTI DD	0	0
	(0)	(0)
2TF2TI DD	0	0
	(0.01)	(0.02)
GI DD	0	0
	(0)	(0)
G2I DD	0	0.01
	(0.04)	(0.09)
G3I DD	0	0
	(0)	(0.01)
D-TFTI DD	0	0
	(0)	(0)
Gamma	0.3**	0.39**
	(0.02)	(0.04)

This table shows the results of a first MLE with all 38 possible strategies. Next we performed a second estimation including only strategies that had weight great than 0.05 in at least one condition. For our final MLE shown in the main text, we then included only the strategies that were present at $p<0.10$ in at least one condition in the second MLE.

Appendix E – Change in conditioning on outcomes over period

Table E1. Cooperation in explicit intentions condition as a function of opponent's actual move (e.g. outcome) in the previous period. Logistic regression with robust standard errors clustered on subject and pairing.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	b/c=1.5, Opponent intended C last period				b/c=4, Opponent intended D last period			
	Last 4 supergames		All supergames		Last 4 supergames		All supergames	
Opponent's Outcome	1.104***	1.666***	-0.0156	0.0786	0.399	0.366	-0.493	-0.456
	(0.289)	(0.556)	(0.474)	(1.186)	(0.337)	(0.433)	(0.520)	(0.492)
Supergame Number		-0.0882	0.0667	-0.0942		0.0584	0.00376	0.0484
		(0.165)	(0.0823)	(0.142)		(0.0942)	(0.0406)	(0.0312)
Period		-0.124		-0.131**		-0.211***		-0.0999*
		(0.100)		(0.0591)		(0.0788)		(0.0515)
Player's intended C last period		3.915***		4.003***		1.237***		1.253***
		(0.466)		(0.467)		(0.464)		(0.269)
Your overall C		6.013***		8.702***		3.326***		3.648***
		(1.010)		(1.147)		(0.781)		(1.135)
Your first period C		-1.209		-2.579***		-1.229***		-1.750**
		(0.837)		(0.837)		(0.428)		(0.806)
Opponent's Outcome X Supergame Number			0.148*	0.227			0.132	0.118
			(0.0883)	(0.174)			(0.0942)	(0.0958)
Constant	1.283***	-3.136*	0.957*	-3.475***	-0.959***	-1.606*	-0.956***	-2.175***
	(0.328)	(1.672)	(0.516)	(0.968)	(0.267)	(0.905)	(0.310)	(0.596)
Observations	795	795	1,708	1,708	293	293	761	761

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table E2. SFEM analyzing first four supergames.

	b/c=1.5			b/c=4		
	No Error	Explicit Intentions	Implicit Intentions	No Error	Explicit Intentions	Implicit Intentions
ALLC	0	0	0.01	0.11†	0.06	0.06†
	(0)	(0)	(0.01)	(0.06)	(0.05)	(0.03)
TFT	0.19*	0.03	0.12*	0.38**	0	0.1**
	(0.09)	(0.03)	(0.05)	(0.09)	(0)	(0.04)
TF2T	0.08†	0	0.05	0.13	0	0.16**
	(0.05)	(0.01)	(0.04)	(0.09)	(0)	(0.06)
TF3T	0	0.03	0.01	0	0	0.05
	(0)	(0.03)	(0.01)	(0)	(0.02)	(0.04)
2TFT	0.04	0	0.11*	0	0	0.03
	(0.05)	(0)	(0.05)	(0)	(0)	(0.03)
2TF2T	0	0	0.11*	0	0	0.15**
	(0.02)	(0)	(0.05)	(0.03)	(0.01)	(0.06)
Grim	0.22**	0	0.07*	0.14*	0	0
	(0.08)	(0)	(0.03)	(0.07)	(0)	(0.01)
Grim2	0.06	0	0.02	0.06	0	0.11**
	(0.04)	(0)	(0.02)	(0.06)	(0.02)	(0.04)
Grim3	0.02	0	0.04	0	0.11	0.07†
	(0.03)	(0)	(0.03)	(0.04)	(0.07)	(0.04)
ALLD	0.20**	0.25**	0.40**	0.09*	0.13*	0.23**
	(0.06)	(0.07)	(0.06)	(0.04)	(0.05)	(0.05)
D-TFT	0.18**	0.06	0.05	0.09*	0.02	0.03
	(0.06)	(0.04)	(0.03)	(0.05)	(0.03)	(0.02)
TFT-I		0.29**			0.34**	
		(0.1)			(0.11)	
TF3T-I		0.06			0	
		(0.04)			(0)	
Grim-I		0.18**			0.19†	
		(0.07)			(0.1)	
Grim2-I		0.03			0.11	
		(0.03)			(0.08)	
TFT-T		0.05			0.04	
		(0.04)			(0.05)	
2TF2T-P		0.04			0	
		(0.05)			(0.04)	
Gamma	*0.43***	*0.43***	*0.57***	*0.37***	*0.51***	*0.51***
	(0.03)	*(0.02)*	*(0.03)*	*(0.03)*	*(0.06)*	*(0.03)*

Appendix F: Forgiveness across conditions

	b/c=1.5			b/c=4		
	No error	Explicit intentions	Implicit intentions	No error	Explicit intentions	Implicit intentions
% Cooperative strategies in SFEM that are forgiving	43%	65%	55%	63%	52%	73%
Descriptive statistics	5%	6%	13%	19%	17%	30%

Considering the SFEM aggregations and forgiveness, it is unclear what the effect of making intentions explicit is: at b/c=1.5, forgiveness is most frequent when intentions are explicit; at b/c=4, forgiveness is least frequent when intentions are explicit.

The results for descriptive statistics are somewhat more consistent. To measure forgiveness using descriptive statistics, we first identify all histories in which (i) at least one subject chose C in the first period, (ii) in at least one previous period, the initially cooperative subject chose C while the other subject chose D and (iii) in the immediately previous period the formerly cooperative subject played D. We then ask how frequently this formerly cooperative subject showed forgiveness by returning to C. We see that in both payoff specifications, forgiveness is similar in the explicit-intentions and the no-error conditions, and lower in the implicit-intentions control.

For Online Publication

Online Appendix A: Decision time analysis regression tables

Table OA1. Predicting cooperation based on decision time. Logistic regression with robust standard errors clustered on subject and pairing.

	(1)	(2)	(3)	(4)
	Last 4 Supergames		All Supergames	
Decision time (log10(sec))	-0.937***		-0.977***	
	(0.214)		(0.168)	
Decision time (sec)		-0.0911***		-0.0991***
		(0.0275)		(0.0223)
b/c	0.418***	0.444***	0.371***	0.388***
	(0.0673)	(0.0677)	(0.0532)	(0.0536)
Supergame	0.0420	0.0383	0.0395***	0.0419***
	(0.0371)	(0.0376)	(0.0118)	(0.0119)
Period	-0.148***	-0.147***	-0.150***	-0.147***
	(0.0143)	(0.0142)	(0.0104)	(0.0103)
Explicit-Intentions	0.141	0.154	0.104	0.109
	(0.257)	(0.258)	(0.195)	(0.195)
Implicit-Intentions	-0.879***	-0.814***	-0.763***	-0.698***
	(0.220)	(0.221)	(0.161)	(0.160)
Constant	0.116	-0.0663	0.227	0.0120
	(0.357)	(0.358)	(0.206)	(0.202)
Observations	11,537	11,604	28,501	28,664
Subject-clusters	338	338	338	338
Pairing-clusters	676	676	1741	1741

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table OA2. Predicting decision time by supergame. Linear regression with robust standard errors clustered on subject and pairing.

	(1) Explicit-Intentions	(2) No-Error	(3) Implicit-Intentions	(4) Implicit-Intentions	(5) Implicit-Intentions	(6) Implicit-Intentions	(7) Implicit-Intentions
	All subjects	All subjects	All subjects	All subjects	All subjects	>67%C Subjects	>67%C Subjects
Supergame	-0.0126***	-0.00998***	0.00672***	0.0163***	0.0191***	-0.000265	0.00261
	(0.00140)	(0.00219)	(0.00169)	(0.00360)	(0.00396)	(0.00192)	(0.00200)
b/c	-0.00536	-0.0576***	0.0240**	0.0414***	0.0368***	0.0294	0.0201
	(0.00782)	(0.00673)	(0.0113)	(0.0132)	(0.0137)	(0.0197)	(0.0195)
Period	-0.0093***	-0.00421***	-0.0114***	-0.00858***	-0.000738	-0.0102***	-0.00274
	(0.00119)	(0.00111)	(0.00139)	(0.00104)	(0.00103)	(0.00168)	(0.00169)
Your Intended Decision	-0.0208	-0.0411***	-0.0923***	-0.0119	-0.00924	0.00537	0.00211
	(0.0144)	(0.0122)	(0.0206)	(0.00765)	(0.00892)	(0.0140)	(0.0124)
Your C Frequency				-0.138***	-0.129**	-0.588***	-0.496**
				(0.0535)	(0.0564)	(0.221)	(0.209)
Supergame X Your C Freq				-0.0199***	-0.0198***		
				(0.00549)	(0.00612)		
Prev Partner's First Move					-0.0239*		-0.0167
					(0.0124)		(0.0125)
Partner's Intention Last Period					-0.0159*		0.00647
					(0.00823)		(0.00962)
Constant	0.374***	0.493***	0.199***	0.160***	0.107***	0.548***	0.439***
	(0.0285)	(0.0267)	(0.0297)	(0.0341)	(0.0392)	(0.142)	(0.136)
Observations	6,552	6,466	15,483	15,483	12,545	4,866	3,949
Subject-clusters	84	92	162	162	162	50	50
Pairing-clusters	404	396	941	941	849	452	410
R-squared	0.061	0.155	0.034	0.072	0.074	0.042	0.026

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Online Appendix B: Experimental Instructions

Main treatments (Explicit intentions)

Instructions:

Thank you for participating in this experiment.

Please read the following instructions carefully. If you have any questions, do not hesitate to ask us. Aside from this, no communication is allowed during the experiment.

This experiment is about decision making. You will be randomly matched with other people in the room. None of you will ever know the identity of the others. Everyone will receive a fixed show-up amount of $10 for participating in the experiment. In addition, you will be able to earn more money based on the decisions you and others make in the experiment. Everything will be paid to you in cash immediately after the experiment.

You will interact numerous times with different people. Based on the choices made by you and the other participants over the course of these interactions, you will receive between $0 and $30, in addition to the $10 show-up amount.

You begin the session with 50 units in your account. Units are then added and/or subtracted to that amount over the course of the session as described below. At the end of the session, the total number of units in your account will be converted into cash at an exchange rate of 30 units = $1.

The Session:

The session is divided into a series of interactions between you and other participants in the room.

In each interaction, you play a random number of rounds with another person. In each round you and the person you are interacting with can choose one of two options. Once the interaction ends, you get randomly re-matched with another person in the room to play another interaction.

The setup will now be explained in more detail.

The round

In each round of the experiment, the same two possible options are available to both you and the other person you interact with: A or B.

The payoffs of the options (in units)

Option	You will get	The other person will get
A:	–2	+8
B:	0	0

If your move is A then you will get –2 units, and the other person will get +8 units.

If you move is B then you will get 0 units, and the other person will get 0 units.

Calculation of your income in each round:

Your income in each round is the sum of two components:
• the number of units you get from the move you played
• the number of units you get from the move played by the other person.

Your round-total income for each possible action by you and the other player is thus

		Other person: A	Other person: B
You	A	+6	-2
	B	+8	0

For example:
If you play A and the other person plays A, you would both get +6 units.
If you play A and the other person plays B, you would get -2 units, and they would get +8 units.
If you play B and the other person plays A, you would get +8 units, and they would get -2 units.
If you play B and the other person plays B, you would both get 0 units.

Your income for each round will be calculated and presented to you on your computer screen.

The total number of units you have at the end of the session will determine how much money you earn, at an exchange rate of 30 units = $1.

Each round you must enter your choice within 30 seconds, or a random choice will be made.

A chance that the your choice is changed

There is a 7/8 probability that the move you choose actually occurs. But with probability 1/8, your move is changed to the opposite of what you picked. That is:

When you choose A, there is a 7/8 chance that you will actually play A, and 1/8 chance that instead you play B. The same is true for the other player.

When you choose B, there is a 7/8 chance that you will actually play B, and 1/8 chance that instead you play A. The same is true for the other player.

Both players are informed of the moves which actually occur, as well as the moves chosen by each player. Thus with 1/8 probability, an error in execution occurs, and you know whether the other person's action was what they chose, or an error.

For example, if you choose A and the other player chooses B then:

- With probability (7/8)*(7/8)=0.766, no changes occur. You will both be told that your move is A and the other person's move is B, and that you chose A and the other person chose B. You will get -2 units, and the other player will get +8 units.

- With probability (7/8)*(1/8)=0.109, the other person's move is changed. You will both be told that your move is A and the other person's move is A, and that you chose A and the other person chose B. You both will get +6 units.

- With probability (1/8)*(7/8)=0.109, your move is changed. You will both be told that your move is B and the other person's move is B, and that you chose A and the other person chose B. You will both get +0 units.

- With probability (1/8)*(1/8)=0.016, both your move and the other person's moves are changed. You will both be told that your move is B and the other person's move is A, and that you chose A and the other person chose B. You will get +8 units and the other person will get -2 units.

Random number of rounds in each interaction

After each round, there is a 7/8 probability of another round, and 1/8 probability that the interaction will end. Successive rounds will occur with probability 7/8 each time, until the interaction ends (with probability 1/8 after each round). Once the interaction ends, you will be randomly re-matched with a different person in the room for another interaction. Each interaction has the same setup. You will play a number of such interactions with different people.

Summary

To summarize, every interaction you have with another person in the experiment includes a random number of rounds. After every round, there is a 7/8 probability of another round. There will be a number of such interactions, and your behavior has no effect on the number of rounds or the number of interactions.

There is a 1/8 probability that the option you choose will not happen and the opposite option occurs instead, and the same is true for the person you interact with. You will be told which moves actually occur, and you will know what move the other person actually chose.

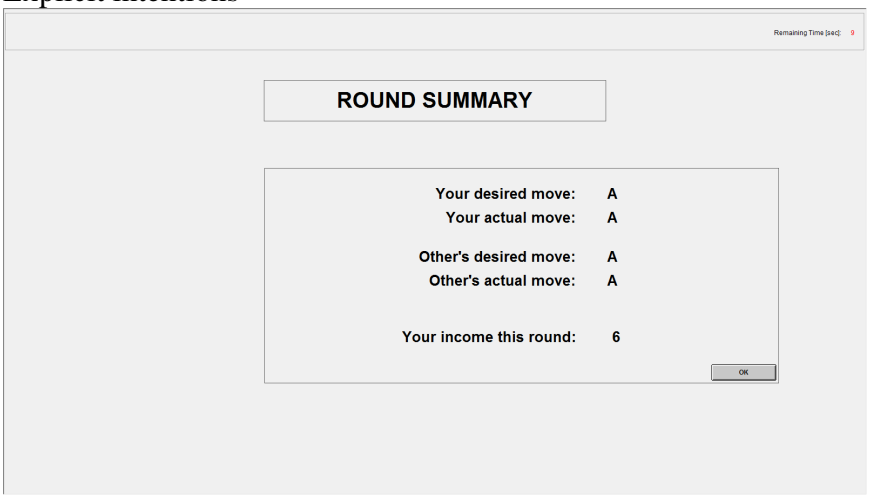
At the beginning of the session, you have 50 units in your account. At the end of the session, you will receive $1 for every 30 units in your account.

You will now take a very short quiz to make sure you understand the setup.

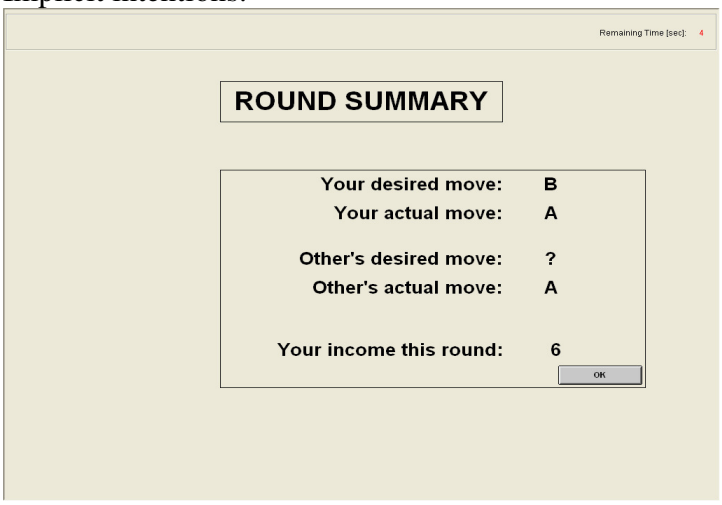
The session will then begin with one practice round. This round will not count towards your final payoff.

Screenshot of the information screen:

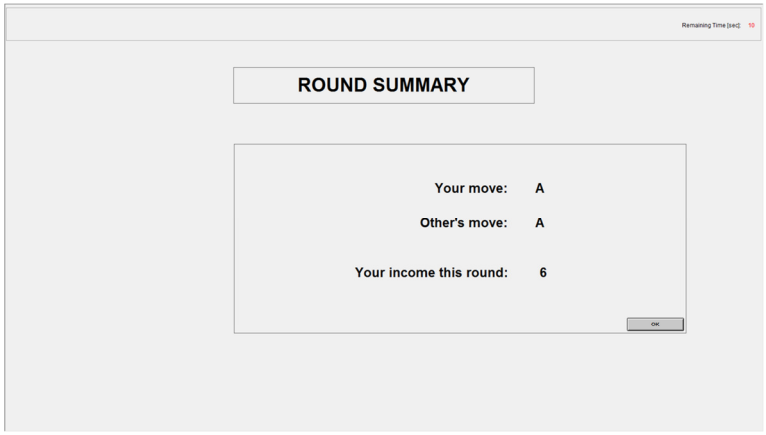
Explicit intentions



Implicit intentions:



No error:



Amazon Mechanical Turk test of the framing of the error term

Error condition:

Imagine you are playing a game with another worker on mTurk. In each round of the experiment, the same two possible options are available to both you and the other person you interact with: A or B.

The payoffs of the options (in units)

Option	You	The other person
A	-2	+8
B	0	0

If your move is A then you will get −2 units, and the other person will get +8 units.

If you move is B then you will get 0 units, and the other person will get 0 units.

Calculation of your income in the game:

Your income in the game is the sum of two components:
• the number of units you get from the move you played
• the number of units you get from the move played by the other person.

A chance that the your choice is changed

There is a 7/8 probability that the move you choose actually occurs. But with probability 1/8, your move is changed to the opposite of what you picked. That is:

When you choose A, there is a 7/8 chance that you will actually play A, and 1/8 chance that instead you play B. The same is true for the other player.

When you choose B, there is a 7/8 chance that you will actually play B, and 1/8 chance that instead you play A. The same is true for the other player.

Both players are informed of the moves which actually occur, as well as the moves chosen by each player. Thus with 1/8 probability, an error in execution occurs, and you know whether the other person's action was what they chose, or an error.

[New page]

Imagine the other player chooses A, and the 1/8 probability switch takes effect, so the other player's choice is changed to B. Therefore, you get 0 cents and the other player gets 0 cents from their action.

To what extent do you think the other person intended to pay 0 cents and give you 0 cents?

Lottery condition:

Imagine you are playing a game with another worker on mTurk. In each round of the experiment, the same two possible options are available to both you and the other person you interact with: A or B.

The payoffs of the options (in units)

Option	7/8 chance	1/8 chance
A	-2 you, +8 other	0 you, 0 other
B	0 you, 0 other	-2 you, +8 other

When you choose A, there is a 7/8 chance that you will get −2 units, and the other person will get +8 units, and 1/8 chance that you will get 0 units, and the other person will get 0 units. The same is true for the other player.

When you choose B, there is a 7/8 chance that you will get 0 units, and the other person will get 0 units, and 1/8 chance that you it iswill get -2 units, and the other person will get +8 units. The same is true for the other player.

Both players are informed of the moves which actually occur, as well as the moves chosen by each player.

Calculation of your income in the game:

Your income in each round is the sum of two components:
• the number of units you get from the move you played
• the number of units you get from the move played by the other person.

[New page]

Imagine the other player chooses A, and the 1/8 probability takes effect. Therefore, you get 0 cents and the other player gets 0 cents from their action.

To what extent do you think the other person intended to pay 0 cents and give you 0 cents?

1 - Did not intend it at all	2	3	4	5	6	7 - Completely intended it
○	○	○	○	○	○	○