619 Methods Section

Greg Stanley

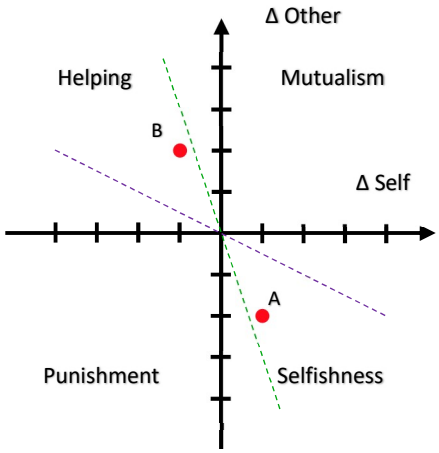
Participants and Apparatus:

The participants were ~100 undergraduates recruited from the University of Michigan Psychology Subject Pool. Participants were run individually in sessions that lasted approximately 1-hour in the lab of Jun Zhang in the basement of the Psychology Department. Participants sat at a desk with a standard laptop with a 10" x 15" monitor and indicated their selections using the mouse. A C++ program displayed the stimuli and saved all responses.

The Game:

The simplest version of the game is where one player is forced to choose between two outcomes with different point values for both players. When compared to the alternative, outcomes can either be good for both players (Mutualism - M), bad for both players (Punishment - P), good for oneself but bad for the other (Selfishness - S), or bad for oneself but good for the other (Help - H).

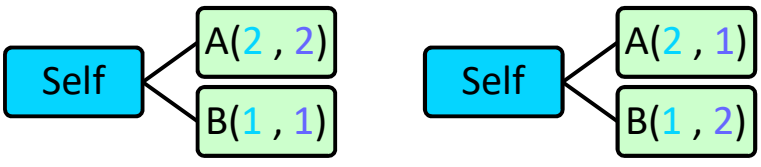
Figure 1: 2D Morality Graph



One player is always active, meaning they control the outcome, while the other player is always passive, meaning they have no control over the outcome. Choices are sequential, not simultaneous, and once made cannot be reversed. No verbal communication is permitted. All games are of perfect information, meaning that all players know the rules, payoffs, possible moves, and know that their opponent knows these things, and knows that their opponent knows that they know these things, etc.

Games are played for points, where 3 is the most preferred outcome and 1 is the least preferred outcome. Each outcome confers each player a specific point value. When comparing two outcomes, players interests are aligned when one outcome is better for both players but are conflicting when either outcome is better for only one of them. By retaining the game structure, while varying the point structure, it becomes possible to generate a wide variety of unique game types that act as experimental conditions.

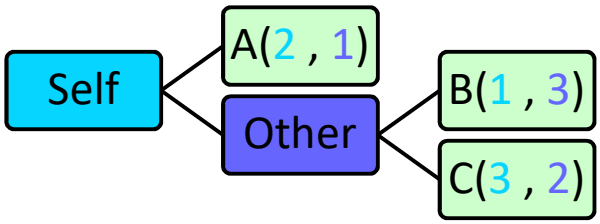
Figure 2: Examples of 1-step games as decision trees



- Left payoffs for player 1 and right payoffs for player 2

Furthermore, this experiment uses a 2-step version of the game, that is identical to the game described above but that includes an extra binary choice for player 2 (as see in Figure 3). In this case, player 1 can either choose to stay or move, where staying ends the game with a certain outcome, but where moving gives control to player 2, who then faces a forced binary choice akin to those described in Figure 2.

Figure 3: Example of 2-step games in decision-tree form



Overview of Experiment:

The goal of is experiment is to establish a new paradigm for testing social decision-making, which can act as a foundation for future research in moral psychology, deception, attribution, recursive Theory of Mind, costly-signaling, trust, and collective action. Future research will also investigate how two individuals establish, build, and break trust over long iterated chains of such games shown in Figure 2, but such complexity will be better understood after first analyzing these games in isolation. The study must demonstrate that it measures what it proports to measure: distinct categories of social actions (M vs P and S vs H). Furthermore, is must demonstrate that these distinct social actions produce divergent judgements, made by observers, of the agent that chooses them. Block 1 will investigate how willing participants are to trust avatars as a result of prior social actions (M vs P and S vs H).

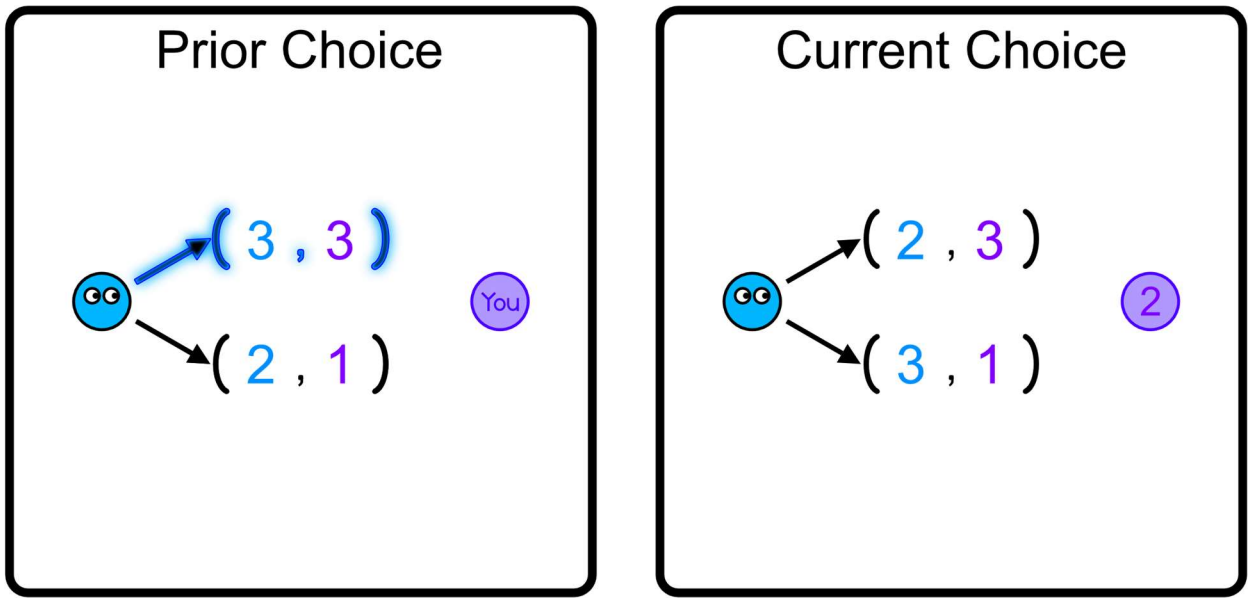
Block 3 is identical to block 1, except that in some conditions, avatars make accidental actions as a result of false beliefs, such as actually choosing S when the avatar believed they were choosing H. Intentionality is integral to moral judgement. Showing that participants willingness to trust avatars depends both on their actual social action *and* their intended social action would demonstrate that this paradigm can capture objective and subjective aspects of social action. This would especially be the case if participants reliably reacted, via trust or distrust, to the intended social action as they would in block 1 to the same overt social action. For instance, a participant may trust a block 3 avatar who intended H but acted S similarly to a block 1 avatar who overtly acted H.

Using a forced-choice matching paradigm, Block 2 will test whether participants can recognize these 4 types of social actions within visual displays devoid of verbal explanation. If participants consistently and successfully match the target game to 1 of 4 templates, animating M, P, H, or S, it will show that participants can generalize these social actions to novel situations. It would take a step toward establishing greater ecological validity and would show that the target display activate participant's social schemas for such things as mutual self-interest (M), punishment (P), altruism (H), and selfishness (S). Furthermore, these templates will include avatars with diverse desires, such as some that prefer lower numbers (1 > 2 > 3), represented in physical features of the templates. If participants correctly match in conditions with diverse desires, it will demonstrate that they are extracting the subjective social essence of the target games, regardless of all physical characteristics of the templates. Thus, blocks 2 and 3 can demonstrate that this paradigm engages participant's reasoning about beliefs and desires, which are two principle components in Theory of Mind.

Block 1:

In the first block, participants interact with a randomized sequence of 216 unique avatars in "games" that resemble social-moral dilemmas. Participants play each avatar in two consecutive games, where the avatar's observed behavior in the first game should influence the participants behavior in the next game. There are 18 types "prior" 1-step games, each with 2 possible choices and there are 6 types of "current" 2-step games, resulting in 216 unique scenarios (18 × 2 × 6), each with a corresponding avatar. These form 216 conditions, where participant's forced binary choice to stay or move in the current game is the dependent variable.

Figure 4: What the screen looks like.



Participants are given the following instructions:

"Act toward each character as you would toward a real person."

- These are hypothetical situations framed as real social encounters.

"Each character is unique. There are 216 characters in block 1."

- The uniqueness of each character is emphasized to reduce the probability that participants will learn general strategies that they can apply across multiple players. Ideally, participants should behave towards each avatar *only* based on their observed prior actions.

"You will have only two consecutive encounters with each character."

- This is their first and only meeting with each avatar.

"You goal is to accrue the highest cumulative payoff across all characters."

- Their goal is self-interested. We do not want participants to use their actions as signals of kindness or cooperation. Participants actions should *only* reflect their evaluation of what the avatar will most likely do next.

"They will act towards you as if they expect to encounter you repeatedly and indefinitely in unknown future situations where the alternatives may or may not resemble past situations."

- Although participants know that they will encounter every avatar only once, participants must understand that the avatars believe that they will meet the participant again and again. The avatars expect these meetings to occur indefinitely but don't know the nature of these situations.

"Their goal is to accrue the highest cumulative payoff across repeated encounters they expect to have with you."

- Understanding this goal + expectation will allow participants to expect the possibility of cooperative behavior from avatars.

The 36 prior game conditions (18-point structures × 2 choices) were drawn from 81 possible point structures (1, 2 or 3 for player 1 × 1, 2 or 3 for player 2)[2]. To reduce the number of conditions and save time, we excluded all point structures for which either alternative conferred an equal payoff for one or both players. Although some of these scenarios are interesting, we have chosen to investigate situations where both players will always have a strong preference for one alternative over the other because they are more likely to be engaging.

Figure 5: All Possible Binary Choices With Discrete Points 1, 2, & 3

	1 1	1 2	1 3	2 1	2 2	2 3	3 1	3 2	3 3
1 1	1 1 1 1	1 1 1 2	1 1 1 3	1 1 2 1	1 1 2 2	1 1 2 3	1 1 3 1	1 1 3 2	1 1 3 3
1 2	1 2 1 1	1 2 1 2	1 2 1 3	1 2 2 1	1 2 2 2	1 2 2 3	1 2 3 1	1 2 3 2	1 2 3 3
1 3	1 3 1 1	1 3 1 2	1 3 1 3	1 3 2 1	1 3 2 2	1 3 2 3	1 3 3 1	1 3 3 2	1 3 3 3
2 1	2 1 1 1	2 1 1 2	2 1 1 3	2 1 2 1	2 1 2 2	2 1 2 3	2 1 3 1	2 1 3 2	2 1 3 3
2 2	2 2 1 1	2 2 1 2	2 2 1 3	2 2 2 1	2 2 2 2	2 2 2 3	2 2 3 1	2 2 3 2	2 2 3 3
2 3	2 3 1 1	2 3 1 2	2 3 1 3	2 3 2 1	2 3 2 2	2 3 2 3	2 3 3 1	2 3 3 2	2 3 3 3
3 1	3 1 1 1	3 1 1 2	3 1 1 3	3 1 2 1	3 1 2 2	3 1 2 3	3 1 3 1	3 1 3 2	3 1 3 3
3 2	3 2 1 1	3 2 1 2	3 2 1 3	3 2 2 1	3 2 2 2	3 2 2 3	3 2 3 1	3 2 3 2	3 2 3 3
3 3	3 3 1 1	3 3 1 2	3 3 1 3	3 3 2 1	3 3 2 2	3 3 2 3	3 3 3 1	3 3 3 2	3 3 3 3

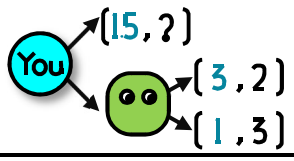
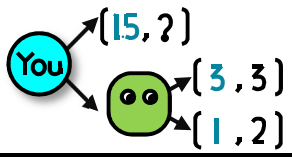
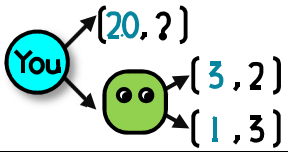
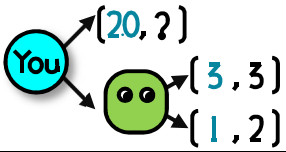
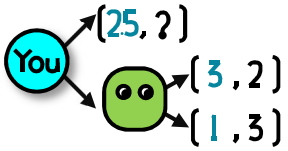
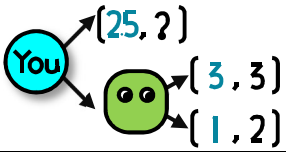
*We selected the 18 payoff structures highlighted in blue.

Moreover, these 36 prior game conditions contain an equal number of all 4 categories of social actions: 8 Mutualism, 8 Punishment, 8 Selfishness, and 8 Help. Additionally, the presence of 3 possible points, instead of only 2, allows us to analyze social actions by the *degree* to which they benefit or harm oneself and the other. For instance, when observing an avatar's prior move, participants may judge a choice to increase oneself by 1-point via decreasing the other by 2-points as extremely selfish because this demonstrates that the avatar cares for his or her interests twice as much as the interests of the other player. However, a choice to increase oneself by 2-points via decreasing the other by 1-point may be more forgivable. Likewise, the presence of 3 possible points allows us to analyze many other aspects of how social actions are perceived. For instance, we can compare two choices resulting in equal relative change in points, yet unequal absolute point values. We can also compare perceptions of categorically different moves, such as M and P, yet that result in the same absolute outcome: choosing (2,2) over (1,1) vs choosing (2,2) over (3,3).

The 6 current game conditions are used to measure different types and degrees of trust. Having observed the avatar's prior choice, the participant will have judged the trustworthiness of this avatar, which can be measured by their decision to stay or move at the first step of the current game. In these two-step games, participants can either choose to stay, thus ending the game and receiving a sure outcome (of 1.5, 2.0, or 2.5 depending on the condition), or to move, thus taking a risk by giving control to the avatar. The avatar now faces a binary choice is either between H (3 for participant, 2 for avatar) and S (1,3) or between M (3,3)

and P (1,2). In either case, the value of the participant's stay outcome will be somewhere between the value of their worst outcome, which is always 1, and their best outcome, which is always 3. Thus, staying = distrust and moving = trust because staying indicates that the participant expects the avatar's choice to result in their worst outcome (1), while moving indicates that the participant expects the avatar's choice to result in their best outcome (3).

Figure 6: All 6 current game types side by side.

	Avatar Choice = H vs S	Avatar Choice = M vs P
Stay = 1.5		
Stay = 2.0		
Stay = 2.5		

As shown in Figure 5, the 6 current game types are composed of two distinct avatar choices and 3 values of the participant's stay option (2 × 3 = 6). The 3 values of the stay option measure degrees of trust. This is because if one's stay option = 1.5, then one has very little to lose by moving if the avatar chooses S or P (0.5), but if one's stay option = 2.5, then one has a lot to lose by moving if the avatar chooses

S or P (1.5). Additionally, 2 types of avatar choices allow different categories of trust to be measured. If it is assumed that participants only ever move in the current games because they are expecting to achieve their highest payoff, then this can either be an expectation that the avatar will move will be H not S or M not P. Trusting someone to sacrifice their interests to your gain (H > S) is different than trusting someone to increase both of your interests (M > P).

Notice that the 2-step version of the game is directly comparable to gambling tasks used to plot utility functions and measure risk attitude. These gambling tasks ask participants to choose between a sure outcome, such as $2, or to flip a coin where tails = $1 and heads = $3, where the sure outcome is always somewhere in between the best and worst outcome. The only difference between such gambling tasks and the 2-step version of the game, is that the second choice is made by an agent, rather than a random coin flip. This single difference means that this paradigm can contrast solo behavior with social behavior and can connect results to a vast literature in behavioral economics.

Potential Problem: We may face a tradeoff between the purity of the experiment and how motivating it is for participants. A pure experiment would give participants no feedback about what avatars choose in the current game if the participant trusts them by moving. An exciting experiment would give this feedback instantaneously by updating their cumulative point total at the top of the screen and this excitement would make participants less likely to respond randomly. In an ideal case, this tradeoff would not exist because participants should treat every avatar independently. However, I'm worried about order effects where participants will learn lessons about who to trust that they will apply to

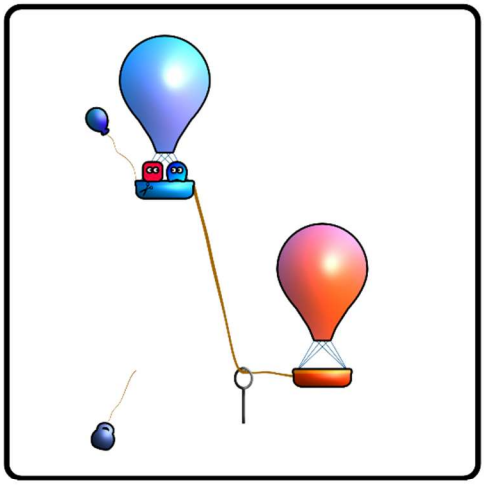
future avatars they haven't met yet. For instance, after avatar x helped but later burned the participant, they might be less likely to trust avatar y. If this feedback is given, I would need to devise a policy for which avatars current choices correspond to their prior choices. A random correspondence would undermine the incentive to carefully assess the trustworthiness of each avatar, while a deterministic correspondence would do the same thing by making it too easy. My current plan is to program avatars with a realistic semi-probabilistic policy for moves in the current games (see Appendix 1). Furthermore, I will prevent participants from seeing these moves and their cumulative point total until after *n* moves. This will still keep them relatively motivated, while mitigating order effects.

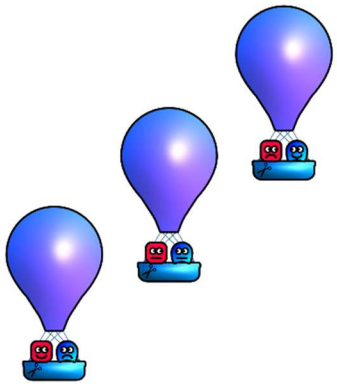
Block 2:

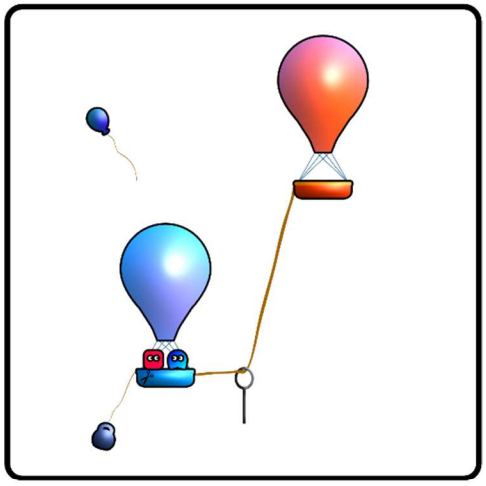
Block 2 is a forced choice matching paradigm where participants attempt to select the template that correctly matches the target. The target and templates are a dynamic between two avatars that the participant observes from the 3rd person. In the center of the screen, the target will be a 1-step game identical to the prior choice games in block 1. In the corners of the screen, will be 4 templates corresponding to actions M, P, H, and S. The templates will be visual representations of these social actions that participants can understand without any verbal instruction or explanation. The templates are images of two hot air balloons tethered to a single pulley, such that one balloon ascending causes the other balloon to descend and vice versa. Before a choice is made, the initial state is for both balloons to be level. A choice is made by using scissors to cut a rope attached to A) a small helium balloon helping the hot air balloon to stay aloft or B)

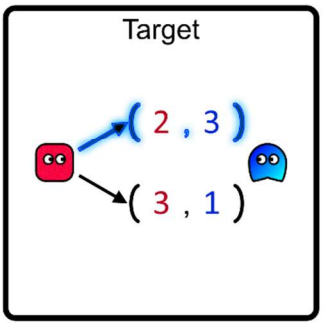
a weight dragging down the hot air balloon. Cutting A causes the hot air balloon to descend and cutting B causes the hot air balloon to ascend. Avatars are either in the same balloon where their interests are aligned (M vs P) or in opposite balloons where their interests are in conflict (H vs S). In every case only one avatar has possession of the scissors and is thus the controlling player. The block begins with a brief simple animation of the hot air balloons beginning level and then descending or ascending as a result of cutting ropes A or B. This animation is followed by the prompt: "There are four templates in each corner of the screen. Select the template that matches the target in the center of the screen."

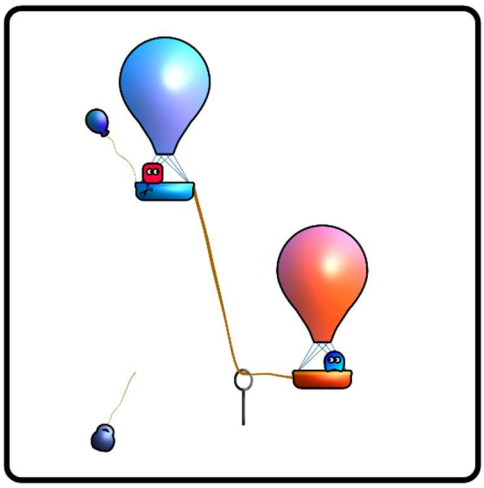
Avatars have diverse preferences for either being high or low. These preferences are not explained but are expressed visually in a picture in the center top of the screen. Here there are three hot air balloons: low, medium, and high. The same two avatars are in each balloon with sad, neutral, or happy faces depending on their preferences for being low, medium, or high. In the beginning of the experiment, participants will see a similar animation of these avatars in a balloon whose expressions change as a function of the height of the balloon. This is done to ensure that participants notice these preferences. Block 2 will introduce 4 new avatars: 2 that prefer being high and 2 that prefer being low. Conditions will include an even number of binary interactions between these players: ↑↑ = ↑↓ = ↓↑ = ↓↓. These conditions will be counterbalanced to ensure that all for players interact equally and equally occupy controlling and passive roles (Player 1 and Player 2). There will be 144 conditions for all 36 targets and 4 diverse desire conditions (36 × 4 = 144).

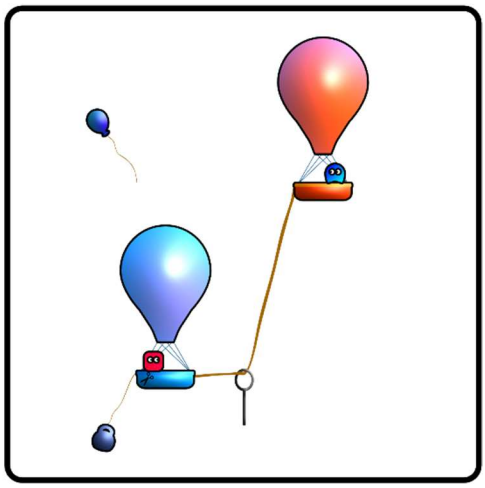







Target
(2 , 3)
(3 , 1)





Block 3:

Block 3 is identical to block 1 with the addition that some avatars have false beliefs about the payoff structure of the prior game. This leads them to make unintentional actions. As shown in Figure 6, avatar's beliefs about the payoff structure are represented in thought bubbles. Just as there are 4 categories of payoff structures, there are 4 categories of beliefs about these payoff structures, creating 16 possible scenarios (4 × 4 = 16).

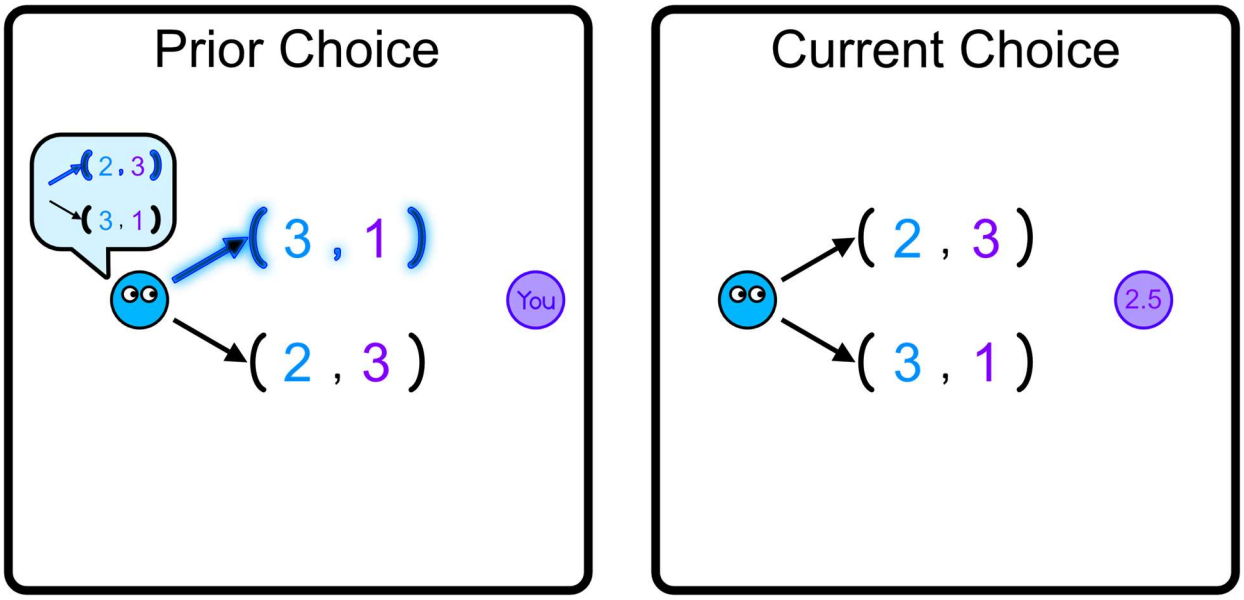
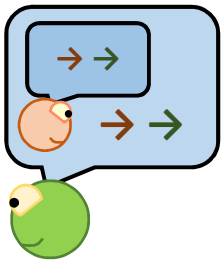
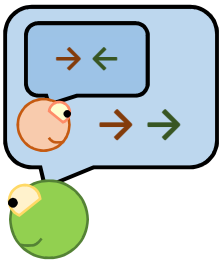
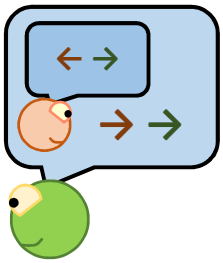
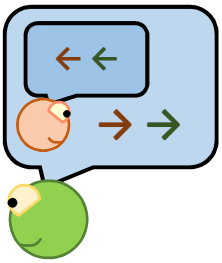
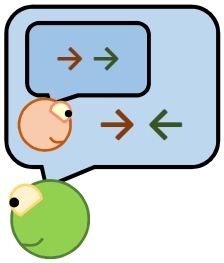
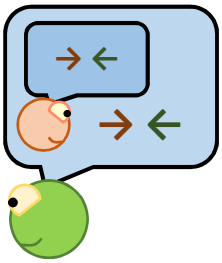
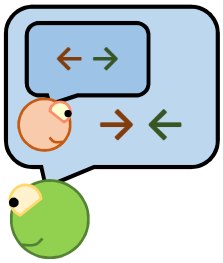
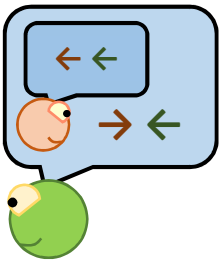
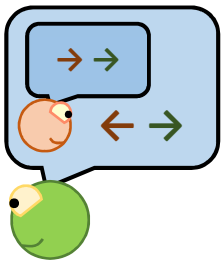
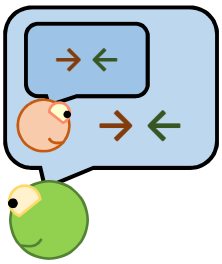
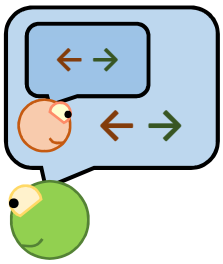
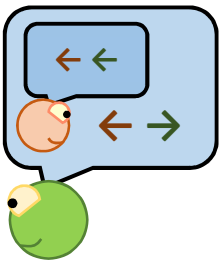
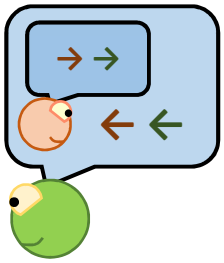
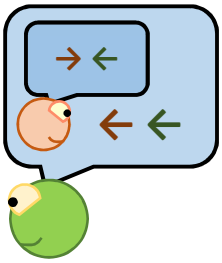
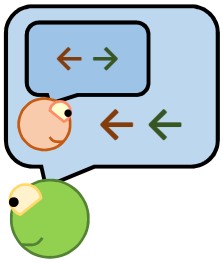
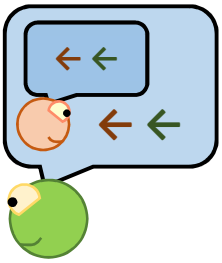


Figure 7: 16 belief-action scenarios

	Belief = M	Belief = H	Belief = S	Belief = P
Action = M				
Action = H				
Action = S				
Action = P				

*These show what the participant (green) believes the avatar (orange) believes.

For every action, there is always one correct belief and 3 false beliefs. In the correct belief scenarios, the believed payoff structure is identical to the actual payoff structure. The 3 false belief scenarios are created by reversing the payoffs of player 1, reversing the payoffs of player 2, or reversing the payoffs of both players. Although it would have been possible to construct scenarios for all 36 prior game types present in block 1, this would take too much time.

Using these 4 belief scenarios multiplies the number of possible conditions by 4, creating 864 possible conditions in block 3. Participants were told that they will never encounter the same avatar again throughout the entire experiment. This is because we want participants to treat each avatar as a clean slate, where their character is judged *only* by their observed actions in the prior game. Thus, between blocks 1 and 3, there are 1080 unique avatars (216 + 864 = 1080) for all 1080 possible conditions.

However, due to time constraints, some participants were not be able to finish all 864 conditions in block 3 and so it was necessary to select specific conditions for them to encounter first. We divided the sequence of conditions into the most valuable first stage comprising 384 conditions and the less valuable second stage comprising 480 conditions. Participants were aloud to continue the experiment until the hour time limit elapsed. Only a minority finished both stages, but a majority (~90%) finished the first stage, and so this is where we focus our analysis. We selected these 384 conditions by categorizing prior game payoff

structures by the relative point gain or loss for both players between both alternatives and then selecting only one payoff structure from each category.

M = (+2,+2) P = (-2,-2)	M = (+2,+1) P = (-2,-1)	M = (+1,+2) P = (-1,-2)	M = (+1,+1) P = (-1,-1)	H = (-2,+2) S = (+2,-2)	H = (-2,+1) S = (+1,-2)	H = (-1,+2) S = (+2,-1)	H = (-1,+1) S = (+1,-1)
3 3 1 1	3 3 1 2	3 3 2 1	3 3 2 2	1 3 3 1	1 3 3 2	2 3 3 1	2 3 3 2
	3 2 1 1	2 3 1 1	3 2 2 1		1 2 3 1	1 3 2 1	1 3 2 2
			2 3 1 2				2 2 3 1
			2 2 1 1				1 2 2 1

*Purple = First Section and Green = Second Section

Thus, the first section of 384 conditions is composed of 8 payoff structures with 2 choices for each making 16 prior game conditions, each with 4 belief conditions and 6 current game conditions (16 × 4 × 6 = 384). The order of these conditions is randomized within the first section, second section, (and the same is true in block 1). Furthermore, the avatars are randomly pared without replacement to each condition at the start of the experiment, meaning that each participant likely encounters a different pairings of avatar to condition. Although the avatars are simple geometric objects of differing colors, we wanted to prevent biases toward specific shapes or colors from distorting the data.

Appendix 1: Current Game Policy

This is the policy for how avatars will behave in the current game depending on their choice in the prior game. As mentioned above, their must be a logical and realistic correspondence between both moves. This policy can neither be random or entirely deterministic because both of these would be disengaging. Instead, the policy will by semi-probabilistic.

Every 1-step game can be plotted as a slope on the 2D Morality Graph. This is the graph where x-axis = Δ Self Interest and y-axis = Δ Other's Interest and the 4 quadrants correspond to M (good for both), P (bad for both), H (bad for self but good for other), and S (good for self but bad for other). The slope is created by plotting a line between the coordinates of both alternatives. Steep negative slopes indicate a dilemma between obligatory (easy) helping and extreme selfishness, whereas gradual negative slopes indicate a dilemma between extreme (difficult) helping and minor selfishness. Thus, the slope of this line + the alternative chosen by the controlling player express information about the character of the controlling player. It shows the minimum extent to which he or she is willing to sacrifice the interests of the other for their gain (or vice versa).

Intuitively, we would expect selfish players to act selfishly next time and helpful players to act helpfully next time. However, following algorithm makes this more precise, by comparing the slope + choice of the prior move to the slope + choice of the current move. *Later I will adapt it such that it can translate between positive slope (M vs P) and negative slopes (H vs S).

m_2 = dilemma, m_1 = first choice

$g = m_2 – m_1$, g = difference between slopes

$z = (g^2)/2$, z = positive g

p = ½z + ½, p = probability of enacting policy below

If $g < 0$ then Help

Elif $g \geq 0$ then Selfish