Moral Values from Simple Game Play

Eunkyung Kim, Ravi Iyer, Jesse Graham, Yu-Han Chang, Rajiv Maheswaran

University of Southern California
Los Angeles, CA 90089
{eunkyung,raviiyer,jesse.graham,yuhan.chang,maheswar}@usc.edu

Abstract. We investigate whether a small digital trace, gathered from simple repeated matrix game play data, can reveal fundamental aspects of a person's sacred values or moral identity. We find correlations that are often counterintuitive on the surface, but are coherent upon deeper analysis. This ability to reveal information about a person's moral identity could be useful in a wide variety of settings.

1 Introduction

It is said that we leave behind a highly revealing digital trail from our myriad online behaviors. We investigate whether a small digital trace, gathered from simple repeated matrix game play data, can reveal fundamental aspects of a person's sacred values or moral identity. In particular, we conduct two studies: in the first, subjects use an online interface to play the "Social Ultimatum Game," a multi-player extension of the well-known ultimatum game [5]; in the second, subjects use a very similar interface to play a different multi-player sequential game. Further details of the games are given in Section 2. In both studies, we find small but significant effects between moral values, such as overall moral identity, Authority, and Fairness, and aspects of the game play, such as the choice of actions and the choice of who to play with in the multi-party game.

Offhand, one might conjure up various stereotypes about values and presumed game play. For example, one might expect that conservatives might be more likely to punish others for unfair actions when a society has already established a norm of fairness, since they typically believe in respect for authority and upholding traditions. However, we show that in fact, liberals' are much more likely to punish, due to their higher degree of desire for fairness. We show that some of these correlations hold across the same studies, under different populations and different game structures. While the effect sizes are not large, they are significant and show promise for further investigation of the degree to which we can extract fundamental aspects of a person's sacred values and personality through observation of simple game behaviors.

2 Background

Related Work. There is a substantial body of work relating observable behaviors to a person's innate personality [12, 17, 16]. When placed in exactly the

same setting, people usually exhibit different behaviors, often correlated with their personality traits. For example, some personality traits can be inferred from inspecting an employee's cubicle [7]. A person's personality not only affects behaviors in the real world but also in the virtual world [13]. A person's concerns, such as fairness and empathy, have also been used to explain behavior in classical game-theoretic matrix games such as Prisoner's Dilemma [18, 2] and the Dictator game [3]. However, these investigations did not measure innate personality traits or moral values and relate those values to the observed game behavior; instead, they typically framed the situation [6] so that the subject would experience empathy, etc. In this paper, we specifically focus on simple repeated games, and examine the relationship between game play and moral values measured using a reliable instrument.

Moral Foundations. Moral values seem to fall within five general categories, or moral foundations [8]. Moral foundations are intuitive sensitivities to particular morally-relevant information. Table 1 shows the five moral foundations.

Moral Foundations	Description
Harm/Care	A concern for caring for and protecting others
Fairness/Reciprocity	A concern for justice and fairness
In-group/Loyalty	A concern with issues of loyalty and self-sacrifice for ones in-group
Authority/Respect	A concern with issues associated with showing respect and obedience to authority
Purity/Sanctity	A concern for purity and sanctity

Table 1. Moral foundations [10]

Based on past research that shows that empathy [15] and people's beliefs about fairness [4] relate to cooperation, we expected that both harm/care and fairness/reciprocity concerns would be significant in a repeated-trials ultimatum game task. We did not expect any relationship to exist due to any of the other three foundations. Past research suggests that political liberals place more emphasis on the harm/care and justice/reciprocity foundations relative to the other three foundations, whereas political conservatives place a relatively equal amount of emphasis on all five foundations [8]. Therefore, the different emphasis people place on these foundations can be used as a proxy for peoples political beliefs. We expected that people who place more emphasis on the harm/care and justice/reciprocity foundations ("liberals") would cooperate more than people who place an equal amount of emphasis on all five foundations ("conservatives"). The 32-item Moral Foundations Questionnaire (MFQ) [9] measures the degree to which people value each of five foundations. Research suggests that the MFQ is highly reliable and valid [9].

The Social Ultimatum Game. This multi-player extension of the classical Ultimatum game was used in the first study. The Social Ultimatum Game models the fact that people operate in societies of multiple agents and repeated pair-wise interactions. These interactions can be thought of as abstract economic transactions that result in surpluses that must be split between the two parties. In each round, we allow each player to propose one transaction with a partner of their choosing, where all transactions result in a $10 surplus that can be split. The proposer must propose a split of the $10. If the other party accepts the proposed split, then the transaction occurs, and the $10 surplus is split as proposed. If the other party rejects the proposed split, then the transaction does not occur, and neither party receives any money. Thus, in each round, a player can make one proposal, and a player can receive between zero and four proposals, and choose to accept or reject each proposal independently. Fig. 1 shows the web interface for the Social Ultimatum Game.

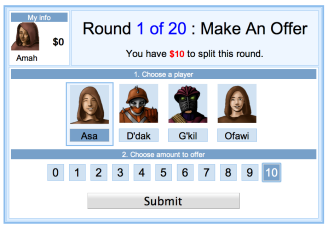

Fig. 1. The Social Ultimatum Game interface. The screen shown is where the player chooses the offer recipient and the offer value.

Trading Game. In the second study, we created another sequential game with a different payoff structure, but which also included notions of fair actions and punishment actions (Fig. 2). In each pair-wise interaction, there is a leader and a follower. The leader can choose either to Trade with or Steal from the follower. In response to a Trade, the follower can choose to complete a Fair or Unfair trade. In response to Stealing, the follower can choose to Punish or Forgive. The rewards intuitively follow the action labels. A leader can ensure a minimum payoff of $10 by choosing to Trade. Or, a leader can take a risk and choose to Steal, hoping that the follower will choose to Forgive, which is in the follower's own self-interest. From a purely economically-rational perspective, this is the equilibrium strategy. The risk is that the leader will receive a negative payoff if the follower decides to act against their own self-interest and Punish instead. As before, we designed a multi-player multi-round game, where in each round, each player can choose a partner and play Trade or Steal with that partner. Thus, in each round, it is possible for a player to receive no Trades or Steals, or up to four Trades or Steals, from the other players. Fig. 2 shows the web interface of the Trading game.

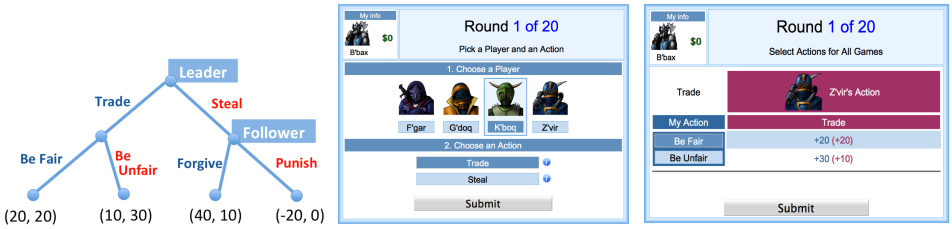

Fig. 2. (Left) Trading Game in extensive form; (Center) User interface for the Trading Game where the leader chooses who to play with, and an action to play; and (Right) Interface where the follower chooses an action.

3 Experiments

We conducted the studies using Amazon Mechanical Turk. In both studies, all participants were invited to finish a compliance test first.This ensures that the experiment subjects understood our game rules and would provide useful data. In the sample game, we showed robot avatars for the agents and gave them screen names such as "Bot-1". After finishing the compliance test, we gave the user a short survey to get background information including gender, occupation, age, education and nationality. Then each participant was given the Moral Foundations Questionnaire. We looked at the timing of question answering and inserted questions with clear correct answers to ensure that the participants were filling out the questionnaire in good faith. In the first study, they would then play a series of four Social Ultimatum games, each time in a different simulated society of four other players. In the second study, they would play a sequence of two Trading games, where we again varied the behavior of the society over time.

In the first study, we focus on two of the societies encountered by the participants. In the "nice" T4T society, all the other agents played tit-for-tat, accepting all offers, and reciprocating whenever possible. In the "harsh" AF7 society, agents used an Adaptive Fairness model which was fit to data produced in an earlier study by an unusual group of subjects who made generous offers, but would not accept offers less than $7 [14].

For the second study, in the first Trading game in the sequence, the other agents in the society would play nicely 80% of the time for the first 10 rounds (playing "Trade" and responding "Fair" or "Forgive"), and not nice 80% of the time for the second 10 rounds (playing "Steal" and responding "Unfair" or "Punish"). In the second Trading game, the other agents would flip this behavior, playing mostly not nice for the first 10 rounds, and nicely the second 10 rounds.

Participants were not told they would play with artificial agents. To simulate that the participants were playing other humans, the avatars and screen names in the actual games were of the same class of that given to the player. We added randomized delays in response time adjusted to match the timings of all-human game play. We paid US$0.50 for all participants and an additional US$0.01 for each $100 earned in the games.

4 Results and Analysis

Study 1. In the first study, we took a more exploratory approach, varying the game types along a number of dimensions and measuring a variety of psychological variables. We were specifically interested in two aspects of game play, the average offer made to other participants and whether individuals chose to spread their offers to many people or to give them to specific others. Our results indicated two significant findings.

One of the most general measures of moral judgment, the Moral Identity Scale [1] interacted with the type of game played, particularly the internal subscale, which measures how central morality is to an individuals self-concept. Specifically, in games where the other players engaged in relatively harsh tactics (AF7), moral identity scores were significantly associated with lower offers ($r = -.20, p < .05$). In contrast, in T4T games, high moral identity scores were associated with higher average offers. This positive relationship was not significant ($r = .11, p = .29$), but the interaction between moral identity scores and game conditions was significant ($F = 3.96, p < .05$). Fig. 3 shows the correlation between Moral Identity Internalization scores and average offers across game conditions. Since tit-for-tat is nominally a fair strategy, this could be taken as indication that individuals who see themselves as more morally motivated reward fair behavior and punish unfair behavior.

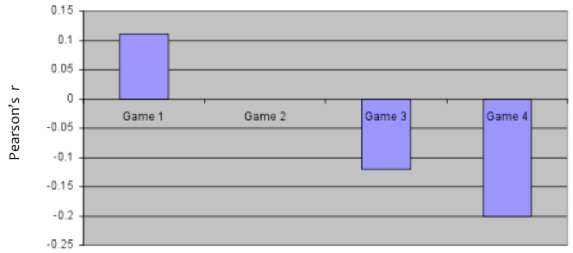

Fig. 3. Correlation of Moral Identity Scores with Average Offers across Games in Study 1: Game 1 is T4T, and Game 4 is AF7. Pearson's r values on Y axis tell us about the strength of the correlation and direction of the association. The closer the correlation value is to either +1 or -1, the stronger the correlation. A positive/negative correlation value means two variables have a positive/negative relationship. If the correlation is 0, there is no association between the two variables.

Many other variables exhibited similar patterns, with moral psychology variables relating to offers and entropy in a step-wise pattern, from "fair" to "unfair" environments though most relationships were not significant. Since our initial approach was exploratory, we did not vary the game environments specifically to create "fair" and "unfair" environments. As such, in Study 2, we sought to explicitly create more definitively "fair" vs. "unfair" game environments. As well, we

noticed that variables tended to cluster in terms of the direction of the correlations, based on whether they were associated with a more liberal or conservative moral profile. For example, conservative values (see [8]) were differentially related to reciprocity and entropy between the T4T game and the AF7 game. As such, we sought to specifically examine whether these values (Ingroup Loyalty, Authority, and Purity) related to aspects of game player behavior differentially, depending on the fairness of the game environment.

We also examined several other behavior characterizations, which attempted to aggregate noisy actions over many rounds. A *window* $w(k, \tau)$ is the set of rounds $\{k, k+1, \ldots, k+\tau-1\}$. Here, we focus in windows involving all 20 rounds and the last 10 rounds. The features for these windows are (1) average offer amount, (2) total score, (3) offer value entropy, (4) offer recipient entropy and (5) reciprocity likelihood. Average offer amount looks at how much each agent offered to the chosen recipient. It is a standard metric for evaluating Ultimatum Game behavior [11]. The total score, i.e., the money made by the participant in the game, is also a standard metric in many economic games of this type [19].

The other metrics try to capture the variance in behavior over time. There may be variability in the offer amounts made by a single player over the course of a game. We introduce the notion of entropy dynamics to capture the changes in this variability over time. Offer amount entropy is a measure of the distribution of offer values over the the window considered. We normalize the standard information theoretic entropy so that the value is bounded above by 1. Similarly, offer recipient entropy is a measure of the distribution of who each player chooses as their offer recipient and is normalized.

Finally, we measure the degree to which players respond to offers by reciprocating with an offer in the next time period. A length-1 reciprocation is when a P_m chooses P_n in round k after P_n made an offer to P_m in round $k-1$, and P_m did *not* make an offer to P_n in round $k-2$. A length-2 reciprocation is when a P_m chooses P_n in round k after P_n made an offer to P_m in round $k-1$, P_m made P_n an offer in round $k-2$, and P_n did *not* make an offer to P_m in round $k-3$. A length-3 reciprocation is defined analagously. Reciprocation likelihood for a particular length is how likely a player engages in such an action given the chance.

The first result is the relationship between authority and recipient entropy. People who thought authority and respect were important tended to explore more in terms of choosing offer recipients. Table 2 shows the effect was stronger in AF7, i.e., the abnormal society, than in T4T, the society where all offers are accepted and reciprocity is high.

A second result was that people who valued in-group (loyalty and self-sacrifice to the group) also showed increased recipient entropy in AF7 (p-values of 0.1027, 0.1332). It seems reasonable that people who value authority and in-group, would find themselves searching more when facing with a society very different from their own. High values on authority and in-group also indicated a higher rate of exploration in terms of offer values with similar significance rates.

		Clustered by recipient entropy	
		windows of 20	windows of 10
Society	T4T	0.1319	0.1144
	AF7	0.0040	0.0698

Table 2. P-values for the differences in authority value between the high/low recipient entropy clusters

Interestingly, harm and fairness, that were initially hypothesized as being potentially key variables, did not seem to have a substantial effect on the measures considered. It is also interesting that for the traditional metrics, offer value and overall score, there were no big differentiators across morality dimensions, but the differences occurred in the temporal metrics.

The third and strongest result is that, again, authority and in-group values, are highly correlated with whether one is in the high and low reciprocation classes. When investigating reciprocation likelihoods, in the cases of length 1,2 and 3, higher authority and in-group values led to lower reciprocation rates. The reciprocation rates difference in the high and low groups were very large (see Table 3) and significant (p-values $\ll 0.001$). Table 4 shows p-values for the differences in authority value and in-group value between the reciprocation high/low clusters.

	Reciprocation likelihood		
	Length-1	Length-2	Length-3
High reciprocation cluster	0.4825	0.6752	0.6707
Low reciprocation cluster	0.2523	0.2064	0.2113

Table 3. Reciprocation rate difference

	Reciprocation likelihood		
	Length-1	Length-2	Length-3
Differences in authority value	0.0040	0.0015	0.0030
Differences in in-group value	0.0453	0.0337	0.1350

Table 4. P-values for the differences in authority and in-group values between the reciprocation high/low clusters

Study 2. Again, we evaluated the entropy measures described earlier, as well as some game-specific features such as: Trade Actions (%): Percentage of choosing Trade action as a proposer; Fair given Trade (%): Percentage of choosing Be Fair action for received Trade games; and Punish given Steal (%): Percentage of choosing Punish action for received Steal games. We calculated the above features for each phase: round 1-10, round 11-20, for each game.

The main result from these experiments is that "liberal" people are more likely to Punish when the leader Steals from them, in relative contrast to "conservative" people who are more likely to Forgive when the leader Steals from them. When the leader plays Steal, then the economically rational reaction by the follower is to Forgive, which provides a payoff of \$10 vs. Punish, which provides \$0 payoffs. However, liberal people tend to give up their own rewards in order to punish what they view as an unfair action by the leader. This can be interpreted as a tendency by liberal people to react more harshly to unfair game play. Fig. 4 shows this relationship between moral liberalness and the choice of Punish actions (p-value = 0.0218), in addition to showing the correlation between moral liberalness and fairness (p-value $\ll$ 0.0001).

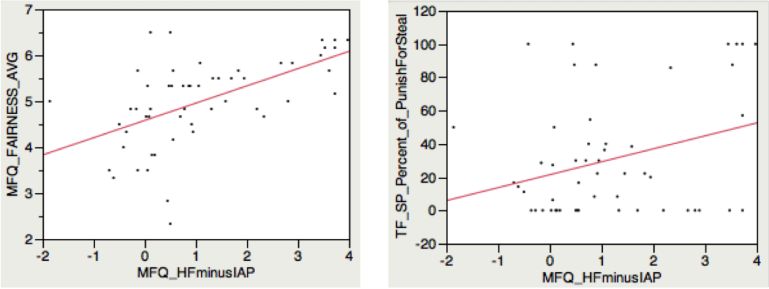

Fig. 4. Positive correlation between fairness and liberal tendency (left), punish action for steal game and liberal tendency (right)

Interestingly, the harm moral foundation, which was initially hypothesized as being a potentially key variable, did not seem to have a substantial effect on the measures considered. When we designed the game, we hypothesized the people with higher harm value on the risk/harm axis would be more likely to choose "Steal", "Be Unfair", "Punish". However, there were no significant effects between a particular subject's harm value and their in-game behaviors.

It is also interesting to note that there were two participants who never chose to Steal as a leader. Also, most of participants either always chose to Forgive, or always chose to Punish, when the leader chose to Steal from them. Finally, we found that there is a correlation between authority and offer recipient entropy, validating the relationship we discovered in Study 1 using the Social Ultimatum Game.

Conclusion. We showed that behavior in simple social games can indicate a player's underlying moral values. While the effect sizes are not large, they are significant and show promise for further investigation of the degree to which we can extract fundamental aspects of a person's sacred values and personality through passive observation of simple behaviors.

Acknowledgements. This material is based upon work supported in part by the AFOSR under Award No. FA9550-10-1-0569 and the ARO under MURI award No. W911NF-11-1-0332.

References

1. K. Aquino and A. Reed. The self-importance of moral identity. *Journal of personality and social psychology*, 83(6):1423, 2002.
2. C. Batson and T. Moran. Empathy-induced altruism in a prisoner's dilemma. *European Journal of Social Psychology*, 29(7):909–924, 1999.
3. G. Bolton, E. Katok, and R. Zwick. Dictator game giving: Rules of fairness versus acts of kindness. *International Journal of Game Theory*, 27(2):269–299, 1998.
4. N. Buchan, R. Croson, E. Johnson, and D. Iacobucci. When do fair beliefs influence bargaining behavior? experimental bargaining in japan and the united states. *Journal of Consumer Research*, 31:181–190, 2004.
5. Y.-H. Chang, T. Levinboim, and R. Maheswaran. The social ultimatum game. In *Decision Making with Imperfect Decision Makers*, 2011.
6. E. Glaeser. Psychology and the market. Technical report, National Bureau of Economic Research, 2004.
7. S. Gosling. *Snoop: What your stuff says about you.* Basic Books, 2009.
8. J. Graham, J. Haidt, and B. Nosek. Liberals and conservatives use different sets of moral foundations. *Journal of Personality and Social Psychology*, 96:1029–1046, 2009.
9. J. Graham, B. A. Nosek, J. Haidt, R. Iyer, S. Koleva, and P. H. Ditto. Mapping the moral domain. *Journal of Personality and Social Psychology*, 101:366–385, 2011.
10. J. Haidt and J. Graham. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20:98–116, 2007.
11. J. Henrich, S. J. Heine, and A. Norenzayan. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83, 2010.
12. J. Jaccard. Predicting social behavior from personality traits. *Journal of Research in Personality*, 7(4):358–367, 1974.
13. L. Jackson, Y. Zhao, E. Witt, H. Fitzgerald, and A. von Eye. Gender, race and morality in the virtual world and its relationship to morality in the real world. *Sex roles*, 60(11):859–869, 2009.
14. E. Kim, L. Chi, Y. Ning, Y.-H. Chang, and R. Maheswaran. Adaptive negotiating agents in dynamic games: Outperforming human behavior in diverse societies. In *Eleventh International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2012.
15. A. Rumble, P. Van Lange, and C. Parks. The benefits of empathy: When empathy may sustain cooperation in social dilemmas. *European Journal of Social Psychology*, 40:856–866, 2010.
16. M. Snyder and W. Ickes. Personality and social behavior. *Handbook of social psychology*, 2:883–947, 1985.
17. E. Staub. *Positive social behavior and morality: I. Social and personal influences.* Academic Press, 1978.
18. J. Tilley. Prisoner's dilemma from a moral point of view. *Theory and decision*, 41(2):187–193, 1996.
19. M. P. Wellman, A. Greenwald, and P. Stone. *Autonomous Bidding Agents: Strategies and Lessons from the Trading Agent Competition.* MIT Press, 2007.