2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)

How Different Identities Affect Cooperation

Wasif Khan
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, CANADA
w23khan@uwaterloo.ca

Jesse Hoey
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, CANADA
jhoey@cs.uwaterloo.ca

Abstract—**Cooperation and competition are a fundamental part of human interaction. In each situation, different people will cooperate or compete with one another in different ways. In this paper, we study the relationship between how people *feel* about the person they are interacting with (the *affective identity* of that person) and their level of cooperation in different circumstances. Standard game-theoretic models provide solutions to what self-interested rational agents would do in various situations. However, humans dont respond rationally in many situations, and the decision to cooperate can be strongly influenced by the identities of the interactants. In this study, over 1,000 participants answered a survey about whether they would cooperate in various framings of the Prisoners Dilemma (PD). These framings are based on the shared cultural sentiments in a three dimensional emotion space (Evaluation, Potency and Activity or EPA) about identities measured in existing large scale surveys. We combined 27 such identities with a set of five different payoff matrices, and provide statistical correlates between the sentiments about identities and likelihood of cooperation. We show that the evaluative (E) dimension is a strong predictor of cooperation, and we discuss the other factors including mixing terms. Our results provide a novel alternative view of cooperation in PD as arising simply from culturally shared sentiments about identities, rather than from payoff estimates.**

1. Introduction

When deciding whether to cooperate with someone, humans consider a wide range of criteria. Most game-theoretic models consider how rational agents would interact with each other and the general conclusion is that they will act in a way that maximizes their payoff. In reality, humans often display a bias towards cooperative behavior, much more so than what is predicted by simple models of rational self-interested agents [1]. These biases are often attributed to inherent properties of people such as fairness, morality or inequity aversion [2], [3], [4], [5], and these properties are added to utility measures to essentially change the nature of the game being played into one that favors cooperative solutions. Many of these properties have emotional interpretations (e.g. fairness is related to guilt), and emotional appraisals have been given rationalistic interpretations as elements in reinforcement learning [6]. However, there are few interpretations of cooperative behaviour that are directly based in a coherent theory of collective emotions. Rather, taking an individualistic approach, most theories relegate emotions to descriptive individual cognitive interpretations of payoff structures or the framing of context. Here we take a different view, and propose that emotional interpretations of a situation are one of the primary motivational forces behind cooperative behaviour. This is in line with neurophysiological evidence of a "low road" guiding action in a way that promotes socio-cultural alignment and homeostasis [7], [8]. We propose that the framing of a situation in terms of identity (i.e. who a person believes they are, what role they are playing, and what role their playing partners are playing) is critical in determining how people will interact with one another. Identity is a well studied notion in psychology, but usually on a denotative level only (e.g. using identity labels such as doctor or mother). Here, we follow an affective social psychological tradition, and view identity as a connotative entity, such that each identity is interpreted in an emotional dimensional space. These emotional interpretations have been proposed as a motivating influence on human choice in social dilemmas [9]. The emotional space is three dimensional, consisting of dimensions of evaluation/valence/pleasure (E/V/P), power/dominance (P/D), and activity (A). This EPA (sometimes called PAD or VAD) space has been extensively studied in sociology [10], later in psychology [11], and has been shown to be cross-culturally very stable and replicable [12], [13].

In this paper, we present results from an experiment in which participants chose preferred actions in a Prisoner's Dilemma (PD) game that was framed in a number of ways according to identity and game payoff matrices. The identities were chosen to span the EPA space, and we show that humans base their decision to cooperate primarily on the identity of the person they are interacting with. Furthermore, we show how influential each of the dimensions in the emotional EPA space are in the decision of whether to cooperate with someone. The main result is that humans base their decision to cooperate primarily on how good the other interactant is (on the Evaluative dimension). We also show that the payoff structure does not have a significant

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

influence on the decision to cooperate.

In Section 2, we review the relevant literature on emotions in cooperative games, the prisoners dilemma, as well as on identity and emotion as related to the study. Then in Section 3, we describe the framing of the PD and the experimental setup. Section 4 describes results followed by a discussion of limitations and future work.

2. Related Work

Rational choice has largely dominated economic theory, leading to inconsistencies when considering simple games with social interdependence (e.g. social dilemmas). Humans in social dilemmas[1] are very good at finding what appear to be non-rational or non-equilibrium solutions that are non-deficient and more globally beneficial. Behavioural economists have tackled this problem by proposing a variety of mechanisms that explain the experimental evidence of prosocial (e.g. cooperative) behaviour in humans. Early work on motivational choice [2] proposed a probabilistic relationship between game outcomes (payoffs) and cooperative behaviour. This led to the proposition that humans make choices based on a modified utility function that includes some reward for fairness [3] or penalty for inequity [4]. More recently, cooperative behaviour has been linked to altruism through factors like kinship, direct reciprocity, or indirect reciprocity via reputation [15]. Further, it appears that fairness or inequity adjustments may not be comprehensive enough to account for human behaviour across all games, and a morality concept that is not based on outcomes can be used as a more parsimonious account [5]. However, the question of how this morality is defined is left as an open question.

Framing effects are a well studied aspect of social dilemmas and economic games, and can be categorized as being either based on *valence* or *context* [31]. While *valence* framings study the relative impact of gains vs. losses, *context* frames study the impact of other aspects of the problem statement, including settings and identities. No proposed theory can explain both types of effects across a range of economic games [31].

Motivational and strategic solution concepts for cooperation that are based on group membership have also been demonstrated [1]. For example, Akerlof and Kranton have proposed an economic model in which an individual's utility function is dependent upon their identity (so called *identity economics*) [16]. Earlier work on *social identity theory* foreshadowed this economic model by noting that simply assigning group membership increases individual cooperation [17], [18]. Other authors have also confirmed that group membership influences individual choice (e.g. [19]). This work has been contested by the counter-argument that it is not the group membership that increases cooperation, but rather that the group membership increases individual's beliefs that others will cooperate (see [1]). The difference is then between group membership as a motivational solution (i.e. being in a group actually changes ones payoff structure in some way), or as a stragtegic solution (i.e. being in a group changes ones beliefs about future events). In recent work, it has been shown that these two solution concepts may not be significantly different. By considering identity as a shared cultural and affective quantity, beliefs about group membership are directly connected to beliefs about strategic choices. That is, the very meaning of the group by definition is an affective one, and this affective sentiment is also explicitly connected to beliefs about behaviours (e.g. good people do good things to good people, but its ok for good people to do bad things to bad people). Using these core principles, and the mathematical structure of affect control theory [20], we have shown that human behaviour in the PD is accounted for more closely [21], [22].

The Prisoner's Dilemma (PD) is perhaps the best studied social dilemma, with early experiments in 1958 [23], [24] leading to the classic experiments of Axelrod in 1981 [25]. Although pure defection is a dominant strategy for any version of the Iterated Prisoner's Dilemma for which the number of rounds (including one) is known by the players [26], human participants do not play the game rationally (in the game theoretic sense), instead showing high rates of cooperation [24]. Human play in the PD is known to deviate from simpler solution concepts like homophily (copying your partner as in tit-for-tat), instead being more in line with moody conditional cooperation, which is more forgiving, and more cooperative in the long term [27].

The unifying theme throughout all this work is that emotional factors strongly influence play and lead to cooperative (seemingly altruistic) behaviour in social dilemmas. Trivers [28] argues that such altruism would almost certainly have been an advantageous trait for early groups of humans to have, citing cases that come at a small cost to the giver, but result in a large benefit for the receiver. Examples include sharing food and tools, helping the wounded, sick, or very young, and sharing knowledge. Further, small, stable groups would have provided ample opportunity for acts of altruism to be applied to kin, or to be reciprocated by the receiver in the future. As a means of encouraging such acts, Trivers proposes the development (or at least co-option) of emotion. Sympathy is an impetus to help those in need; gratitude promotes returning the favour and guilt dissuades from cheating others in the group. Antos et al. [29] have worked from these concepts and focussed on trust, showing that humans trust agents whose emotions match their actions. Van Kleef proposes that emotional signals will carry different meanings in cooperative vs. competitive situations, but leaves open the question of how cooperativeness is appraised in the first place [30].

The connotative (affective) meaning of a person's identity forms the basis for a well-established sociological view of human interaction based on emotional alignment called Affect Control Theory (ACT) [20]. The idea is that people try to establish an affective meaning for each partner in an interaction, and then use a non-linear dynamical system

[1] A social dilemma is a game with *uncompensated interdependencies* (externalities) [1]: each person's actions in the game affect other persons without their explicit consent (e.g. without compensating them).

to make consistent predictions about future actions. The affective meanings and the dynamical system are culturally shared and so result in consistent behaviour in culturally similar partners. Affective meanings are defined in a three dimensional emotional space of Evaluation - goodness versus badness, Potency - powerfulness versus powerlessness and Activity - liveliness versus torpidity. These three dimensions were uncovered in large cross-cultural surveys using semantic diferentials in the 1950s [10], and more recently replicated [13]. The scales for E,P and A are defined (by historical convention) between -4.3 and $+4.3$. For example, an individual can hold the identity "child" which maps to the EPA point $[1.45, -0.76, 2.10]$ as most people regard a child as someone moderately good and active but not very powerful. The identity "scrooge" maps to the EPA point $[-1.36, 0.08, -1.62]$ as a scrooge is someone relatively bad and passive but neutral on potency. ACT states that an individual will seek to act in ways that conform to their identity. For example, a scrooge may "scoff at" $[-1.62, -0.90, -1.00]$ someone as this behavior would maintain the identity of the scrooge [20].

Sociologists have compiled numerous large-scale surveys of affective meanings of identities, behaviours (action words), modifiers (e.g. emotions) and settings (e.g. locations and institutions). These surveys are carried out using semantic differential scales which ask respondents to rate concepts on scales with opposing adjectives at each end (e.g. good $\leftrightarrow$ bad, or strong $\leftrightarrow$ weak). Surveys have been carried out in the USA, Canada, Germany, and Japan, with recent work in North Africa. Surveys provide mean ratings by both men and women (although these are rarely significantly different). The survey methodology and some of the datasets are described in [33]. All EPA ratings in this paper come from the Ontario 2001 dataset [32].

In this work, we demonstrate a strong correlation between cooperation in the PD with a person's evaluation (on a good vs. bad axis) of the *identity* of their partner. This finding is in line with the morality concept of Capraro and Rand [5] which shows that more cooperation is induced, the more "good" the partner is evaluated to be. Our work is also well aligned with our previous studies of the PD, showing that human behaviour is well replicated by emotional identity dynamics [22]. Our study connects evaluation scores for identites gathered in large-scale sociological surveys in the affect control theory literature from 2001, with cooperative behaviour in the PD in our study in 2017. This arises because of the temporal stability of culturally shared sentiments about identities, as previously established [33].

3. Prisoner's Dilemma

In the classic framing of the Prisoner's Dilemma, you and a fellow gang member are caught at a crime scene and the police take you to separate rooms. They have evidence to convict each of you for a crime that serves a one year prison sentence. The principal crime has a three year prison sentence but the police have insufficient evidence to convict either of you for that - unless there is a verbal testimony against the other person. The police offer each of you the option to testify against the fellow gang member and in return, they will let you off the one year sentence and go free. The only catch is - if both of you testify against each other, then both are found guilty and will serve a reduced two year prison sentence for the principal crime. This is represented (see Table 1) with T = Temptation = 0, C = Cooperation = -1, D = Defection = -2 and S = Sucker = -3. For this to be considered a Prisoner's Dilemma, we need $T > C > D > S$ as is the case in this example. $C > D$ implies mutual cooperation is superior to mutual defection while $T > C$ and $D > S$ imply that defection is the dominant strategy for two rational agents.

TABLE 1. General and example reward matrix. Payoffs are for (row,column) players.

	Cooperate	Defect
Cooperate	C,C	S,T
Defect	T,S	D,D

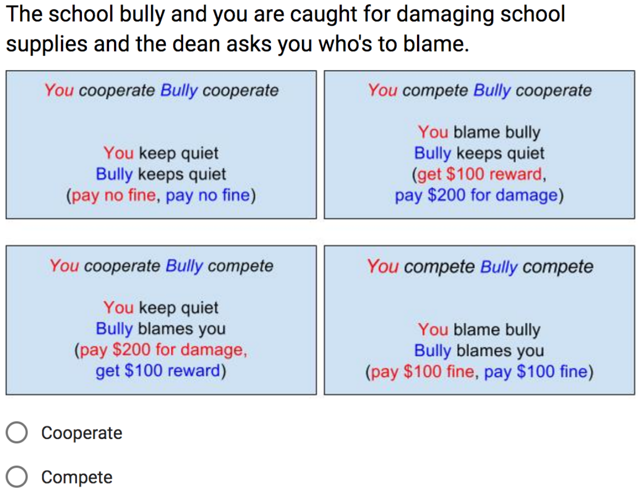
	Silent	Betray
Silent	-1,-1	-3,0
Betray	0,-3	-2,-2

When considering this through the cultural affective lens, the identity "gang member" maps to the EPA point $[-1.46, 0.78, 1.07]$. Most people would opt for betraying the gang member as this is the reasonable option (on a socio-emotional level) when considering where "gang member" is located in the EPA space (E very negative). However if we consider the identity "neighbour" which has a rating of EPA=$[1.35, 0.08, 0.04]$, we hypothesize that a significantly larger percentage of individuals would opt for cooperation.

To test this, we chose 27 identities spaced evenly in the central part of the Evaluation-Potency-Activity (EPA) space, taken from $E = \{1.5, 0, -1.5\} \times P = \{1.5, 0, -1.5\} \times A = \{1.5, 0, -1.5\}$. For each combination of E,P, and A in these sets, we sought an identity in the Ontario 2001 ACT lexicon [32] that was close[2]. To test the influence of various identities in the Prisoner's Dilemma, we kept the payoff matrix ratios approximately constant but altered the quantities to see if a gain or loss influences the decision. The payoff matrices presented range from $T > C > D > S > 0$ - only positive outcomes whether you cooperate or compete, to $0 >= T > C > D > S$ - only non-positive outcomes. The payoff matrices can be sorted according to how positively they frame the PD, using a discretization of $K = D/MAX(ABS(T, C, D, S))$ into five payoff matrix *categories*, $c = \{1, 2, 3, 4, 5\}$, bounded by $K \geq 0.5, 0.5 > K > 0, K \equiv 0, 0 > K \geq -0.5, -0.5 > K$, respectively. These categories order the PDs according to how positively vs. negatively the payoffs are framed, with larger category index c indicating a more negative framing.

From the 5 payoff matrices and 27 identities - we constructed 27 Prisoner's Dilemmas assigning payoff structures to identities at random. These PDs have different context (identity) and valence (gains/losses) framings. For example, the PD presented in Figure 1 is for identity "bully"

[2] We used either male or female ratings depending on which was closest to the target point.



The school bully and you are caught for damaging school supplies and the dean asks you who's to blame.

You cooperate Bully cooperate You keep quiet Bully keeps quiet (pay no fine, pay no fine)	*You compete Bully cooperate* You blame bully Bully keeps quiet (get $100 reward, pay $200 for damage)
You cooperate Bully compete You keep quiet Bully blames you (pay $200 for damage, get $100 reward)	*You compete Bully compete* You blame bully Bully blames you (pay $100 fine, pay $100 fine)

○ Cooperate

○ Compete

Figure 1. Example question from the survey for the identity "bully"

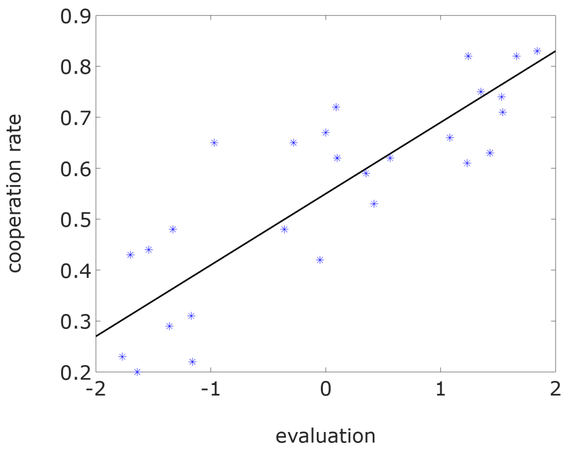


Figure 2. Cooperation rate vs. Evaluation showing the best fit line.

(EPA=$[-1.77, 1.17, 1.29]$: negative, powerful and active) with a payoff structure of $T = 100, C = 0, D = -100, S = -200$ (defection is framed as a loss).

The study was conducted online, after receiving full clearance from the University of Waterloo Research Ethics Board. Subjects were recruited through a list of online contacts maintained by the first author, and are mostly made up of younger north american adults. The online questionnaire consisting of the 27 PDs were sent out to $1,194$ participants. Participants were 95% males from the USA (45%), Canada (26%) and Germany (8%), in age groups 18-24 (53%), 25-34 (23%) and 35-44 (14%). For each PD, users were asked whether they would cooperate or compete (defect) with the identity they were interacting with. PDs were presented in a randomized order for each participant. Due to the length of the questionnaire, many participants did not complete it entirely resulting in just over $1,000$ responses per PD. As the order was randomized, the missing data was randomly spread across the PD frames.

4. Results

The results of the questionnaire, ordered by cooperation rate, are presented in Table 2. The first column shows the number of participants who rated that PD. Then, each row shows the label used in the PD framing as well as the E,P, and A values and which gender did the rating, along with the payoff matrix category and the T, C, D and S values and units for those values. The rightmost column shows the cooperation rate across all subjects. Time-based units are shown as negative (as they are lost time).

With the results of the questionnaire above, we can see over 50% of participants opted for cooperation in 17 of the 27 scenarios, while a self-interested rational agent would never cooperate in all 27 scenarios. We also see there are 7 scenarios where over 70% of humans would cooperate and 3 scenarios over 80%. There is clearly a strong relationship between the cooperation rate and the evaluation dimension, highlighted in Figure 2. Thus, how "good" the partner is in the frame is the strongest predictor of how likely participants were to cooperate with them. χ^2 tests for the main effects of the sign of each identity rating (E,P, and A either > 0 or $<= 0$), as well as for the sign of the defection payoff (D) show that while Evaluation and Potency are highly significant ($p < 0.001$), Activity is not and the defection payoff is only weakly significant (see Table 3). Thus, while the valence framing shows an effect, the context framing in terms of evaluation of the identity is much stronger. Power (P) also shows an effect, but is reversed (cooperation is more likely with weaker identities).

A multivariate linear regression with all six degree-2 polynomial features in E,P and A yields the parameters shown in Table 4, with a coefficient of determination of $R^2 = 80\%$, and an overall p-value of 0.000012.

This confirms our result that the Evaluation dimension has the biggest influence on cooperation, specifically a positive 14% influence to cooperation as compared to the next biggest influence of 3%.

The above result shows us a clear relationship between the Evaluation dimension and decision to cooperate. We also investigated how influential the payoff structure was in comparison by dividing the data into the five payoff categories as described in Section 3. We found no consistent or significant effects in cooperation rates, but the strong correlation between Evaluation and cooperation was still present in each category.

We augmented the linear regression with the payoff category $\in \{1, 2, 3, 4, 5\}$, and another for the product of category with evaluation. These parameter had small coefficients of -0.005 and 0.004 ($t = -0.29$ and $t = 0.28$, $Pr(> |t|) = 0.78$ and 0.79, respectively), suggesting that the payoffs have little impact on the cooperation rate, possibly because they don't present inequity or fairness differences.

TABLE 2. COOPERATION RATE FOR THE 27 IDENTITIES IN THE STUDY, SORTED FROM HIGHEST TO LOWEST.

	identity					Payoff						cooperation
N	label	E	P	A	rater	category	T	C	D	S	unit	rate
1034	father	1.84	1.78	0.02	male	5	0	-1000	-2000	-3000	dollars	0.83
1039	brother	1.66	1.28	1.35	male	2	3000	2000	1000	0	dollars	0.82
1017	father-in-law	1.24	1.37	-1.15	female	4	1000	0	-1000	-2000	dollars	0.82
1044	neighbor	1.35	0.08	0.04	male	5	-1	-2	-4	-5	hours	0.75
1027	fisherman	1.53	0.20	-1.40	female	1	400	300	200	100	fish	0.74
1043	mistress	0.09	-0.15	1.88	male	5	0	-1000	-2000	-3000	dollars	0.72
1021	janitor	1.54	-1.25	-1.82	female	5	0	-1000	-2000	-3000	dollars	0.71
1028	atheist	0.00	-0.20	0.35	male	3	2000	1000	0	-1000	dollars	0.67
1035	trainee	1.08	-1.36	1.38	female	2	3000	2000	1000	0	dollars	0.66
1034	auditor	-0.97	1.48	-0.79	male	2	3000	2000	1000	0	dollars	0.65
1019	stoner	-0.28	-1.16	0.23	female	1	90	80	70	65	% grade	0.65
1011	babysitter	1.43	-0.04	1.58	male	3	2000	1000	0	-1000	dollars	0.63
1042	crony	0.10	-0.08	-0.75	male	5	0	-1000	-2000	-3000	dollars	0.62
1018	cripple	0.56	-1.86	-1.29	female	1	2500	2000	1250	0	dollars	0.62
1034	autistic person	1.23	-1.14	0.20	female	3	2000	1000	0	-1000	dollars	0.61
1038	shrink	0.35	1.42	-1.19	male	3	2000	1000	0	-1000	dollars	0.59
1027	attorney	0.42	1.71	0.40	male	3	2000	1000	0	-1000	dollars	0.53
1032	runaway	-0.36	-1.42	1.56	male	3	2000	1000	0	-1000	dollars	0.48
1022	tramp	-1.33	-1.48	1.24	female	5	0	-1	-2	-3	months in jail	0.48
1024	slacker	-1.54	-1.29	0.34	female	1	2500	2000	1250	0	dollars	0.44
1032	hoodlum	-1.70	-0.15	1.68	male	4	1000	0	-1000	-2000	dollars	0.43
1048	spy	-0.05	1.28	1.31	male	5	0	-1	-2	-3	weeks in jail	0.42
1029	degenerate	-1.17	-1.46	-1.26	female	5	0	-1	-3	-4	months in jail	0.31
1037	scrooge	-1.36	0.08	-1.62	male	1	2500	2000	1250	0	dollars	0.29
1023	bully	-1.77	1.17	1.29	male	4	100	0	-100	-200	dollars	0.23
1034	antagonist	-1.16	1.06	0.30	female	4	1000	0	-1000	-2000	dollars	0.22
1041	scoundrel	-1.64	0.04	-0.04	male	3	2000	1000	0	-1000	dollars	0.20

TABLE 3. EFFECTS OF BINARIZED VARIABLES. N^+, N^-: NUMBER OF VALUES $> 0, <= 0$, RESP., C^+/C^-: COOPERATION RATES.

	N^+	N^-	C^+	C^-	χ^2
Evaluation	15	12	0.69	0.40	2244.2 ***
Potency	14	13	0.55	0.58	24.6 ***
Activity	17	10	0.57	0.56	1.6
Defection payoff	12	15	0.57	0.55	6.9 *

TABLE 4. LINEAR REGRESSION PARAMETERS AND SIGNIFICANCES. TARGET VARIABLE IS DIFFERENCE IN COOPERATION RATE FROM 0.5.

	Coeff.	Est Std.Error	t value	$Pr(>\|t\|)$
(Intercept)	0.049	0.020	2.442	0.024 ·
Evaluation	0.14	0.017	7.835	1.61e-07 ***
Potency	-0.011	0.018	-0.604	0.552
Activity	0.00061	0.017	0.036	0.972
EP	0.028	0.017	1.655	0.114
EA	-0.018	0.015	-1.245	0.227
PA	-0.024	0.016	-1.487	0.153

5. Limitations and Future Work

A clear limitation of the study is that over 95% of the participants were males, over 70% were North American, and over 50% were between the age of 18-24, creating a sampling bias. Further, the identity ratings shown in Table 2 are also from a sample of North American undergraduates in 2001. Given the known individualism of North Americans [34], the results may not generalize to more collectivistic cultures, and we hypothesize these results are skewed towards competition when compared globally. Further, linear regression gives an R^2 value of 80%, suggesting there may be better statistical models, or more relevant factors we have not accounted for.

A major assumption we make is to assume only the identity and payoff struture influences the decision to cooperate. We could also map the setting (i.e. location of the event) into EPA space, using affective settings ratings from the same body of research [35], and this may affect the conclusions. We did not account for the different payoff values in Table 2, as it was difficult to find relationships bewteeen the units. Nevertheless, the payoff structures were similar in terms of inequity, and so the results do not inform us about the relative importance of identity frames vs. outcome structures beyond differences in gains/losses.

Nowak provides us with a theoretical framework addressing how cooperation could evolve through natural selection [15]. Although we don't investigate the matter, we hypothesize there is a relationship between the EPA space and the various ways cooperation can evolve. For example, kin selection is a method that allows cooperation to evolve. Looking at the relationship between cooperation and identities such as "brother" and "daughter" among other kin related identities would yield some interesting results.

The data suggests that the potency dimension may have a negative correlation to cooperation, and that while Evaluation and Activity alone may be positively correlated to cooperation, the mixed effect of Evaluation and Activity seems negatively correlated to cooperation. The data also suggest that the more negative the payoffs were - the more people were inclined to defect, possibly a reflection of the risk averse nature of humans [14]. These suggestions could be further analyzed. The 55% intercept could be investigated to see if this is an indication of a bias towards cooperation.

In the future, we plan to investigate other payoff structures, social and moral dilemmas, iterated games and networked games. The connections to behavioural economics are of interest, and we plan to investigate further how utility and emotions are connected in the human mind.

6. Conclusion

Taking an socio-affective view of cooperation, we derive 27 framings of the Prisoner's Dilemmas and sample over 1,000 people asking them whether they would cooperate or compete in various scenarios. We provide descriptive results on how humans would respond in various situations as opposed to normative results by considering how rational agents would respond. Using multivariate linear regression we show the decision to cooperate is strongly influenced by the identity in the frame. Our results point to an alternative explanation for human behaviour in social dilemmas as arising from shared cultural and affective interpretations of social situations, rather than from payoff structures alone.

Acknowledgment

This work was supported in the framework of the Trans-Atlantic Platform by NSERC and SSHRC. Thanks to Tobias Schröder for discussions and to our reviewers.

References

[1] P. Kollock, "Social dilemmas: the anatomy of cooperation," *Annual Review of Sociology*, vol. 24, pp. 183–214, 1998.

[2] D. M. Messick and C. G. McClintock, "Motivational bases of choice in experimental games," *Journal of Experimental Social Psychology*, vol. 4, pp. 1–25, 1968.

[3] M. Rabin, "A theory of fairness, competition and cooperation," *The American Economic Review*, vol. 83, no. 5, pp. 1281–1302, 1993.

[4] E. Fehr and K. M. Schmidt, "A theory of fairness, competition, and cooperation," *The Quarterly Journal of Economics*, vol. 114, no. 3, pp. 817–868, 1999.

[5] V. Capraro and D. G. Rand, "Do the right thing: Preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality," May 2017, SSRN.

[6] T. M. Moerland, J. Broekens, and C. M. Jonker, "Emotion in reinforcement learning agents and robots: A survey," *Machine Learning*, 2017, accepted.

[7] A. R. Damasio, *Descartes' error: Emotion, reason, and the human brain*. Putnam's sons, 1994.

[8] J. Zhu and P. Thagard, "Emotion and action," *Philosophical Psychology*, vol. 15, no. 1, pp. 19–36, 2002.

[9] N. Asghar and J. Hoey, "Monte-Carlo planning for socially aligned agents using Bayesian affect control theory," in *Proc. Uncertainty in Artificial Intelligence (UAI)*, 2015, pp. 72–81.

[10] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The Measurement of Meaning*. Urbana: University of Illinois Press, 1957.

[11] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[12] C. E. Osgood, W. H. May, and M. S. Miron, *Cross-Cultural Universals of Affective Meaning*. University of Illinois Press, 1975.

[13] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional." *Psychological Science*, vol. 18, pp. 1050 – 1057, 2007.

[14] A. Tversky and D. Kahneman, "The framing of decisions and the psychology of choice," *Science*, vol. 211, no. 4481, p. 453458, 1981.

[15] M. A. Nowak, "Five rules for the evolution of cooperation," *Science*, vol. 314, p. 15601563, 2006.

[16] G. A. Akerlof and R. E. Kranton, "Economics and identity," *The Quarterly Journal of Economics*, vol. 115, no. 3, pp. 715–753, 2000.

[17] H. Tajfel and J. C. Turner, "An integrative theory of intergroup conflict," in *The social psychology of intergroup relations*, S. Worchel and W. Austin, Eds. Monterey, CA: Brooks/Cole, 1979.

[18] M. A. Hogg, "Social identity theory," in *Contemporary Social Psychological Theories*, P. J. Burke, Ed. Stanford University Press, 2006, ch. 6, pp. 111–136.

[19] G. Charness, L. Rigotti, and A. Rustichini, "Individual behavior and group membership," *The American Economic Review*, vol. 97, no. 4, pp. pp. 1340–1352, 2007.

[20] D. R. Heise, *Expressive Order: Confirming Sentiments in Social Actions*. Springer, 2007.

[21] J. D. Jung and J. Hoey, "Grounding social interaction with affective intelligence," in *Proceedings of the Canadian Conference on AI*, Victoria, BC, 2016.

[22] ——, "Socio-affective agents as models of human behaviour in the networked prisoner's dilemma," 2017, http://arxiv.org/abs/1701.09112.

[23] M. M. Flood, "Some experimental games," *Management Science*, vol. 5, no. 1, pp. 5–26, 1958.

[24] M. Deutsch, "Trust and suspicion," *Journal of conflict resolution*, pp. 265–279, 1958.

[25] R. Axelrod and W. Hamilton, "The evolution of cooperation," *Science*, vol. 211, no. 4489, pp. 1390–1396, 1981.

[26] D. M. Kreps, P. Milgrom, J. Roberts, and R. Wilson, "Rational cooperation in the finitely repeated prisoners dilemma," *Journal of Economic Theory*, vol. 27, pp. 245–252, 1982.

[27] J. Grujić, C. Gracia-Lázaro, M. Milinski, D. Semmann, A. Traulsen, J. A. Cuesta, Y. Moreno, and A. Sánchez, "A comparative analysis of spatial prisoner's dilemma experiments: Conditional cooperation and payoff irrelevance," *Scientific reports*, vol. 4, p. 4615, 2014.

[28] R. L. Trivers, "The evolution of reciprocal altruism," *Quarterly Review of Biology*, vol. 46, p. 3557, March 1971.

[29] D. Antos, C. D. Melo, J. Gratch, and B. J. Grosz, "The influence of emotion expression on perceptions of trustworthiness in negotiation," in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[30] G. A. V. Kleef, "An interpersonal appraoch to emotion in social decision making: The emotions as social information (easi)," in *Advances in Experimental Social Psychology*, M. Zanna, Ed. New York: Academic Press, 2010, vol. 42, pp. 45–96.

[31] P. Gerlach and B. Jaeger, "Another frame, another game?" in *Proceedings of Norms, Actions, Games*, A. Hopfenspitz and E. Lori, Eds. Toulouse: Toulouse: Institute for Advanced Studies, 2016.

[32] N. J. MacKinnon, "Mean affective ratings of 2,294 concepts by Guelph university undergraduates, Ontario, Canada in 2001-3," computer file, 2006.

[33] D. R. Heise, *Surveying Cultures: Discovering Shared Conceptions and Sentiments*. Wiley, 2010.

[34] D. Schwalb, K. Murata, and B. Schwalb, "Cooperation, competition, individualism and the inter-personalism in japanese fifth and eighth grade boys," *International Journal of Psychology*, vol. 24, no. 1, pp. 617–630, 1989.

[35] L. Smith-Lovin, "The affective control of events within settings," *Journal of Mathematical Sociology*, vol. 13, pp. 71–101, 1987.