The Good, the Bad, and the Unflinchingly Selfish: Cooperative Decision-Making can be Predicted with high Accuracy when using only Three Behavioral Types

ZIV EPSTEIN, Pomona College
ALEXANDER PEYSAKHOVICH, Yale University
DAVID G. RAND, Yale University

The human willingness to pay costs to benefit anonymous others is often explained by social preferences: rather than only valuing their own material payoff, people also care in some fashion about the outcomes of others. But how successful is this concept of outcome-based social preferences for actually predicting out-of-sample behavior? We investigate this question by having 1067 human subjects each make 20 cooperation decisions, and using machine learning to predict their last 5 choices based on their first 15. We find that decisions can be predicted with high accuracy by models that include outcome-based features and allow for heterogeneity across individuals in baseline cooperativeness and the weights placed on the outcome-based features (AUC=0.89). It is not necessary, however, to have a fully heterogeneous model – excellent predictive power (AUC=0.88) is achieved by a model that allows three different sets of baseline cooperativeness and feature weights (i.e. three behavioral types), defined based on the participant's cooperation frequency in the 15 training trials: those who cooperated at least half the time, those who cooperated less than half the time, and those who never cooperated. Finally, we provide evidence that this inclination to cooperate cannot be well proxied by other personality/morality survey measures or demographics, and thus is a natural kind (or "cooperative phenotype").

Additional Key Words and Phrases: Machine learning, social preferences, behavioral economics, prosociality

1. INTRODUCTION

The willingness to pay costs to help others is a key feature of human behavior, and is essential for the success of human societies. Understanding why, when, and to what extent people engage in this "cooperative" behavior is thus a major focus of research across the social and biological sciences. Numerous mechanisms have been proposed which demonstrate how cooperation can actually be in one's long-run self-interest: for example, forces such as repetition [Fudenberg and Maskin 1986], reputation [Nowak and Sigmund 1998], and dynamic social networks [Rand et al. 2011] can compensate the cost of cooperating today with benefits received tomorrow (for a review, see [Rand and Nowak 2013]).

While much of human cooperation can be explained by these mechanisms, there is no question that people sometimes cooperate even when it is clearly not in their long-run self-interest to do so (e.g. in anonymous interactions with strangers). Given the

This work is supported by the John Templeton Foundation and the VIA Institute. See the Acknowledgements section before REFERENCES.
Author's addresses: Z. Epstein, Computer Science Department, Pomona College; email: ziv.epstein@pomona.edu; A. Peysakhovich, Human Cooperation Lab, Yale University. Currently at Facebook News Feed; email: alex.peys@gmail.com, D. G. Rand, Human Cooperation Lab, Department of Psychology, Department of Economics, School of Management, Institute for Network Science, Yale University ; email: david.rand@yale.edu.

EC'16, July 24–28, 2016, Maastricht, The Netherlands.
http://dx.doi.org/10.1145/2940716.2940761

absence of strategic reasons to cooperate, efforts to explain this latter "altruistic" form of cooperation have focused on psychological motivations.[1]

Attempts to formalize non-strategic motivations for cooperation, and to assess the predictive power of these motivations, often take the form of incorporating "social preferences" (also called "other-regarding" preferences) in economic models of decision-making. The psychological benefits of increasing efficiency [Becker 1976; Charness and Rabin 2002] or decreasing inequity [Bolton and Ockenfels 2000; Fehr and Schmidt 1999] may outweigh the material costs of cooperating in 1-shot anonymous settings; and so, social preference models can successfully predict cooperation when a model based on self-interest alone would predict none.

Here, we use machine learning to shed new light on social preferences. In recent years, machine learning has emerged as a powerful tool for generating accurate predictions. Machine learning has been widely embraced outside the academy by businesses, organizations, and governments interested in predicting a wide range of phenomena from advertisement click-thrus to purchasing decisions to violent crime, as well as by academic geneticists and neuroscientists. This approach, however, has gained little traction among behavioral scientists, who often see machine learning as a black-box tool that does not generate meaningful insights, just accurate predictions. In this paper, we show how machine learning can be useful for basic social science research.

We do so by generating a large dataset of 21,340 incentivized cooperation decisions made by 1,067 human participants, and using machine learning to answer the following questions regarding social preferences. First, we ask how well, quantitatively, an extremely general model of outcome-dependent social preferences (i.e. preferences that take into account the other person's payoff) can predict decisions. Critically, characterizing cooperative behavior using social preference models is only useful insomuch as it allows one to quantitatively predict *future* behavior. Otherwise, these models are just a restatement of one's data – if every dataset generates a new social preference model, the model is not shedding light on the underlying psychology of altruism, but instead is just adding extra degrees of freedom for fitting data. Thus we in particular ask the question of how well outcome-based preferences can do at *out-of-sample* prediction, which is essential for evaluating the usefulness of the social preference approach. Furthermore, we consider the accuracy of predicting individual cooperation decisions, rather than just ability to recreate general patterns across experiments. (Importantly, we explore social preferences that drive unilateral behavior in the absence of strategic motives for giving, unlike the work of Wright & Leyton-Brown 2012,2014 who use machine learning to investigate strategic reasoning about the behavior of others in multilateral games.)

Second, we characterize the extent of *heterogeneity* in social preferences.[2] There is clear evidence of large and stable differences between individuals in their cooperativeness [Peysakhovich et al. 2014; Yamagishi et al. 2013]. Yet when social preference models are applied, it is often assumed that all people have identical preferences (the "representative agent" assumption, e.g. [Charness and Rabin 2002]). Furthermore, when this assumption is relaxed, it is typically done in an ad hoc way. For example, various papers have allowed 2 different types [Fehr et al. 2004, 2007], 4 different types [Fehr and Schmidt 1999], or have estimated separate utility functions for every participant [Andreoni and Miller 2002; Blanco et al. 2011; Fisman et al. 2007]. To explore the role

[1] It is important to differentiate between this proximate psychology, which is not based on long-run payoff-maximization, and the ultimate evolutionary or strategic forces that may have given rise to this psychology [Bear and Rand 2016; Persakhovich and Rand 2015].

[2] Peysakhovich & Naecker (2015) take a similar approach to explore models of choice under risk and ambiguity.

of heterogeneity across individuals, we first compare out-of-sample prediction accuracy for the representative agent approach (which ignores individual variation and focuses on game payoffs) and a person-focused model which ignores game payoffs and just characterizes variance across individuals in cooperativeness. We then evaluate models that consider both, allowing individual heterogeneity in responsiveness to payoffs (i.e. heterogeneous preferences), and ask how accuracy varies as a function of the number of types allowed: Is it necessary to go all the way to the extreme of unique preferences for every participant in order to achieve high predictive power? Or can an intermediate number of types yield similar accuracy?

Finally, we ask how well cooperative type can be proxied by a wide range of survey measures and demographics commonly used in economics and psychology (and interactions between these items). Can a participant's type be predicted with reasonable accuracy using these other measures or is cooperativeness a "natural kind", that is, a "separate" personality trait of cooperativeness? Distinguishing between these possibilities has important implications for theorizing about prosociality and its origins and cognitive implementation.

We also make a methodological contribution: [Peysakhovich and Naecker 2015] compare the predictive power of machine learning to economic models of decision-making in the case of risk and ambiguity. They argue that because machine learning is optimized heavily for prediction these models can be thought of as the upper bound of 'explainable variance' in a task. We note that if this number is low it suggests that the data set is not very useful (either because the task itself is too complicated, the participants are noisy or we are not recording the relevant features). We find high 'explainable variance' in our data set and this further demonstrates that cooperation games on Mechanical Turk are a powerful tool for behavioral science researchers [Horton et al. 2011].

2. METHODS

To create a dataset of cooperation decisions, we recruited 1067 U.S. residents (M_{age}= 35.6, 55.6% female) from the online labor market Amazon Mechanical Turk (MTurk) [Horton et al. 2011].[3] Consistent with standard payment rates on MTurk, participants received a show-up fee of $\$0.75$, and then had the opportunity to earn additional money based on their decisions in the study. In particular, each participant made a series of 20 randomly generated binary-choice decisions between money allocations for themselves and another anonymous MTurk worker (the recipient).[4] For each decision, the participant chose between

(1) An option that was better for themselves, in which the participant received x cents and the recipient received y cents (the "selfish" option); or
(2) An option that was better for the recipient, in which the participant received $x - c$ cents and the recipient received $y + c(f + 1)$ (the "cooperative" option).

Thus, choosing the cooperative option entailed paying a cost c to give a benefit of $c(f+1)$ to the recipient. The selfish option payoffs x and y for each decision were integer values randomly sampled from a uniform distribution over the interval $[0, 50]$; the cost was randomly sampled from a uniform distribution over the interval $[0, x]$; and the efficiency factor of cooperation f was randomly sampled from a uniform distribution over the interval $[0, 4]$. Participants were informed at the outset that of their 20 decisions,

[3] As per common practice on MTurk, we excluded additional participants who did not finish the study, failed any of a series of attention check questions, or had duplicate MTurk IDs or IP addresses.

[4] Our procedure is similar to that of [Charness and Rabin 2002], but uses randomly generated (rather than pre-specified) payoffs for each choice, and confronts each participant with a unique set of choices.

one would be selected at random for actual payment. The study took participants an average of 13 minutes, and they earned an average of $1.01 (including a $0.75 show-up fee).

After making their cooperation decisions, participants completed a post-experiment questionnaire with a suite of survey measures and demographics that may be relevant to cooperative behavior: gender, age, income level, education level, political party affiliation, fiscal and social conservativism, belief in God, extent of prior experience with MTurk studies (as in [Rand et al. 2014]) and skepticism about whether there actually was a real recipient; non-incentivized measures of risk aversion, ambiguity aversion, competitiveness, and intertemporal choice; and a one item measure of generalized trust (as in [Rand et al. 2012]), one item measures from the faith in intuition and need for cognition scales [Epstein et al. 1996] (as used in [Rand and Kraft-Todd 2014]), the short form measure of the "Big Five" personality domains [Gosling et al. 2003], a 24-item measure of character strengths (adapted from [Peterson and Seligman Peterson and Seligman]) classified into a 4-factor structure of prosociality, self-determination, intellectualism, and self-control, and a set of two "trolley problems" about the permissibility of harming one person to save multiple others [Greene et al. 2001].

To assess the ability of outcome-based preferences to predict cooperative behavior, we trained logistic ridge regression models[5] on the cooperative choice data, using participants' first 15 decisions as the training set and the last 5 decisions as the hold-out set.[6]

We used the following feature set: x, y, c, f, an $f > 0$ indicator (indicating that cooperating is non-zero-sum), the quantity $[((x-c)-(y+c(f+1)))-(x-y)]$ (the change in inequality caused by cooperating), and an indicator for whether this change in inequality is positive; as well as squared versions of each of these terms (except the indicators), and all two-way interactions. Note that the set of two-way interactions includes the term cf, which is the efficiency gain created by cooperation. Thus this feature set nests, among other things, the standard measures of inequity and efficiency.

We chose the regularization parameter by using a 5 fold cross-validation within the training set with each fold of the cross-validation including 3 decisions of each participant. We implemented all the analysis using the `R` package *glmnet* [Friedman et al. 2010].

To explore heterogeneity in preferences, we compared the predictive power of a model using just these features (such that the same set of coefficients was applied to all participants – the representative agent assumption) to models in which these features interacted with varying numbers of dummies indicating participant type (such that the coefficients could vary with type), as described in more detail below. There is a simple Bayesian interpretation to this "add all features + interactions" approach – we are training individual-level (ot type-level) models, but we are pooling data from all individuals and regularizing each model towards the average.

To ask how well cooperative type can be proxied for by other measures, we predicted participant type with logistic ridge regression using as features the responses to the

[5]A key concept in machine learning is the "bias-variance tradeoff." Models which are more flexible fit data better (so have lower bias) but they are more sensitive to the input data (have higher variance) and so can overfit in sample. One way to make the bias variance tradeoff is to regularize - fit complex models to the input data but penalize the model for its complexity. Ridge regression [Hoerl and Kennard 1970] performs this regularization by starting with a standard linear regression model, allowing for complexity in the function mapping inputs to outcomes (either by adding more features or by adding basis expansions such as polynomials) but penalizing the model for including larger coefficients.

[6]Using the first 5 decisions as the hold-out set does not qualitatively alter our results.

survey measures and demographics collected in the post-experiment questionnaire as well as their interactions.

All of these various analyses involve assessing how successfully a given model predicts choices. When doing so, it is essential to account for skew in the dataset: if exactly 50% of the decisions were cooperation and 50% selfishness, it would be sufficient to straightforwardly ask what fraction of decisions were correctly predicted. However, this is not the case: only 17.8% of the decisions in our dataset were cooperative. Therefore, we used the standard approach of calculating the AUC ("area under the receiver operating characteristic (ROC) curve" [Bradley 1997]), which provides a balanced measure of accuracy. AUC captures the likelihood that, when faced with a selfish choice and a cooperation choice, the model correctly predicts which is which. Thus an AUC of 0.5 indicates no predictive power, while an AUC of 1 indicates perfect prediction.

3. RESULTS

We began by assessing the predictive power of a model making the representative agent assumption. This ridge regression included all of the outcome-based features, but no participant-specific dummies. This model therefore forces each feature to have the same coefficient for each participant. We found that this representative agent model had non-negligible predictive power, generating an AUC of 0.69.

Next we assessed a model at the other extreme, which ignored the payoffs for any given decision and only considered individual heterogeneity in average giving rates. This ridge regression included dummies for each participant, but no outcome-based features. Interestingly, this model performed much better than the representative agent model, yielding an AUC of 0.83. We found that a large fraction of this explanatory power comes from being able to discern extremely selfish types (those who never cooperate) from those who cooperate at least sometimes - a model with just a single "giver vs non-giver" dummy generated an AUC of 0.71.

Performance was improved even more by using a fully heterogeneous social preference model, in which every participant was allowed to have their own unique set of social preferences (i.e. coefficients for the outcome-based features). This ridge regression included the outcome-based features, dummies for each participant, and interactions of these dummies with each of the outcome-based features. This fully heterogeneous model yielded an AUC of 0.89. Thus, at least in the context of our particular experimental design, both the payoff details of the decision and the person making the decision mattered; within-person regularities matters substantially more than the payoffs; and the interaction between the two provides the best predictive power.

But is it necessary to go to the extreme of full heterogeneity in preferences in order to capture these individual differences? To explore this issue, we considered an additional set of models with intermediate amounts of heterogeneity. In these models, we allowed either two, three, or four behavioral types, where a participant's type was defined by how many of their 15 training observations involved them choosing the cooperative option.[7] Figure 1 shows the extent of variation in our dataset of cooperativeness during the training observations.

The precise definitions of each type were set in a principled way so as to maximize predictive power, as follows. In models that allowed for two different types, we clas-

[7]A potential issue with this classification is that the extent to which a given person cooperated in the first 15 trials is influenced not only by that person's inclination to cooperate, but also by the particular set of randomly generated payoffs that person faced - for example, people who faced choice sets with higher efficiency gains (i.e. larger f values) are likely to have cooperated more. Importantly, however, this variation in the payoffs of games 1 thru 15 should, if anything, work against us by degrading the model's predictive power, as all participants faced the same distribution of choices over the last 5 games.

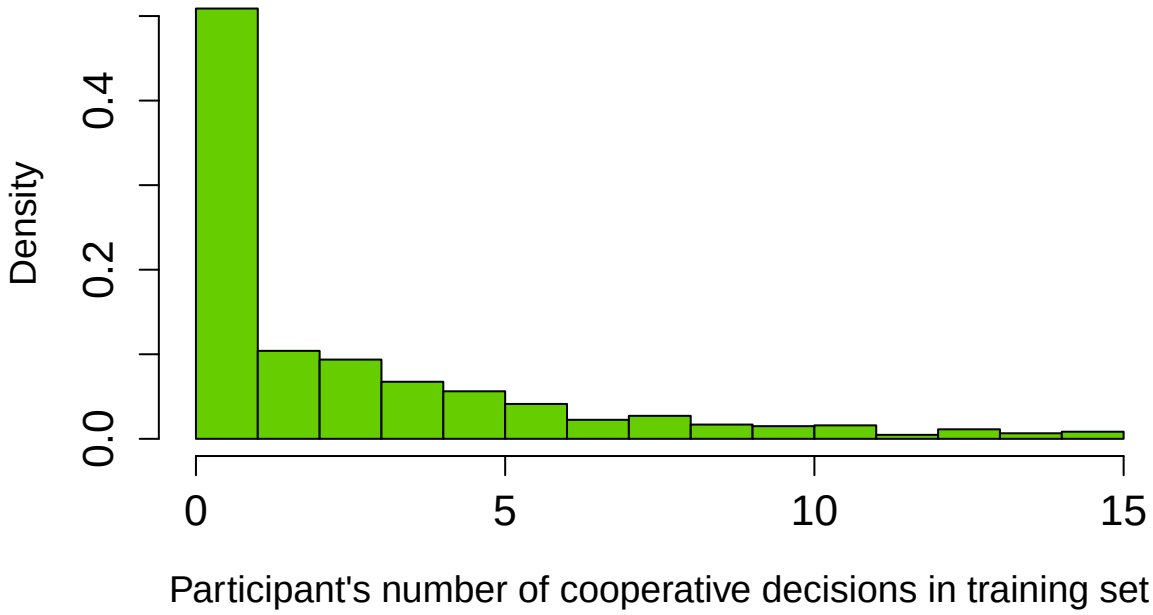

Fig. 1. Distribution of participant cooperativeness during the 15 training observations.

sified participants as "less cooperative" if they cooperated in fewer than k of their 15 training observations and "more cooperative" if they cooperated in k or more decisions. For each possible value of k (0 to 15), we then ran a ridge regression in which a "cooperated at least k times" dummy was interacted with each outcome-based feature ($k = 0$ classifies all participants the same way, and thus is equivalent to the representative agent formulation). The resulting AUCs are shown in Figure 2, and can be used to select the optimal 2-type definition (i.e. the definition that had the most predictive power). The best-performing type definition was $k = 2$, and yielded an AUC of 0.83. This AUC is much better than the representative agent model, but still substantially lower than the fully heterogeneous model.

Next we considered 3-type models. In these models, we classified subjects as "least cooperative" if they cooperated in fewer than j of their 15 training observations; as "intermediately cooperative" if they cooperated at in at least j but fewer than k of their 15 training observations; and as "most cooperative" if they cooperated in k or more of their 15 training decisions. As for the 2-type models, we then iterated over all possible values of j and k to find the pair that maximized the predictive power of the model in out of sample prediction. The resulting AUCs are shown in Figure 3. The best-performing type definition was $[j = 1, k = 6]$, and yielded an AUC of 0.88. This AUC is quite close to AUC generated using the fully heterogeneous model.

We then considered 4-type models with an analogous procedure, with cooperation thresholds i, j, and k. The best performing 4-type definition was $[i = 1, j = 4, k = 9]$,

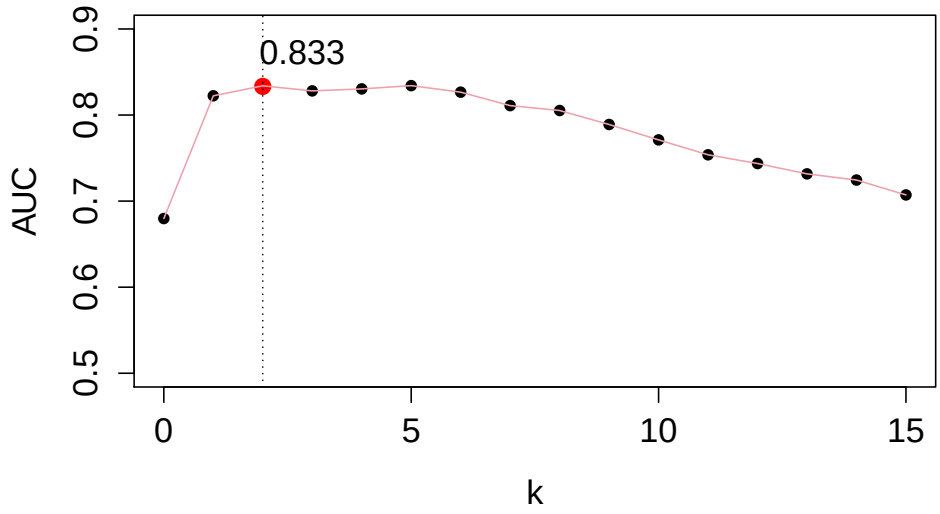

Fig. 2. Predictive power of 2-type models that include a"more cooperative" dummy indicating whether participants cooperated in k or more of their 15 trial observation (and interactions between this dummy and all outcome-based features), as a function of k.

and yielded an AUC of 0.88. As this performance is not better than the 3-type model, we concluded that three behavioral types most effectively captures individual variation in outcome-based preferences. For ease of comparison, the AUC of the best-performing model of each type is shown in Figure 4.

We complemented this analysis of AUCs produced by 2-, 3-, and 4-type models with a bottom up approach to determining the number of types. To do this, we trained the fully heterogeneous model above and took the implied regression coefficients for each individual. We then represented each individual as their vector of coefficients. We then performed standard k-means clustering on the vector of coefficients. Note that changing a coefficient by a single unit also changes the predictions of the model for the same input by one unit. Thus, this analysis allows us to check whether a small number of types (represented by centroids in their cluster) can meaningfully summarize the heterogeneity in utility functions.

As shown in Figure 5, we found a large improvement in variance explained when increasing from 2 clusters to 3 clusters, but substantially less improvement when increasing from 3 clusters to 4 (or above). Thus the clustering analysis provides further support for a small number of types being able to explain a large fraction of individual heterogeneity.

Finally, we asked how well cooperative type can be predicted using survey measures and demographics. To do so, we classified participants' type based on the best-performing 3-type model:

(1) *least cooperative*: those who cooperated in fewer than one training observations (N=414);
(2) *intermediately cooperative*: those who cooperated in at least one but fewer than seven training observations (N=492); and

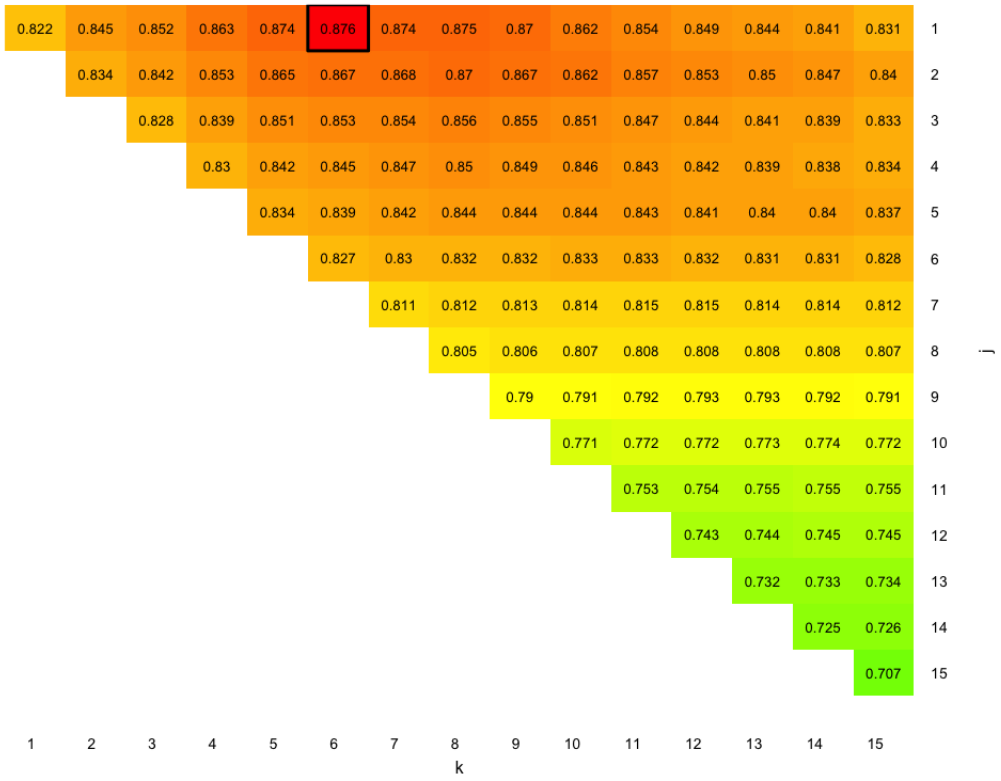

Fig. 3. Predictive power of $3-$type models that include types that cooperate at least $0-i$ times, $i-j$ times or $j-15$ times in the training set for various values of i and j.

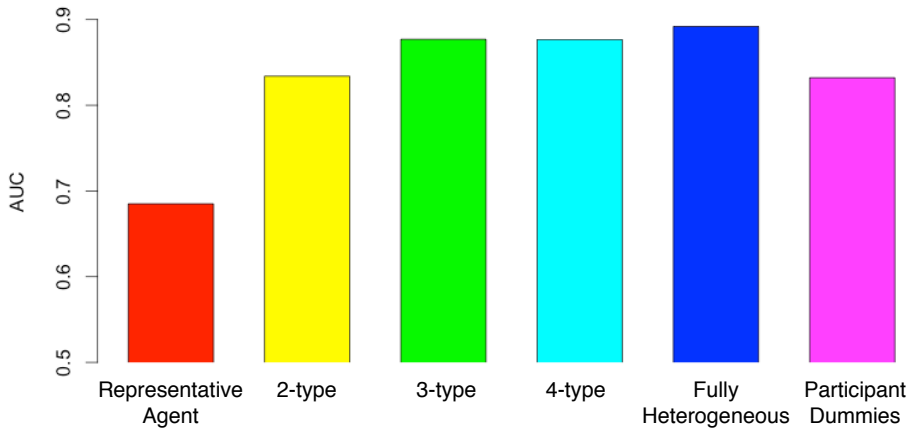

Fig. 4. Predictive power of cooperative decision making for the representative agent model, the best-performing 2-type, 3-type and 4-type models, the fully heterogeneous model, and the model containing only participant dummies (and no information about the payoffs).

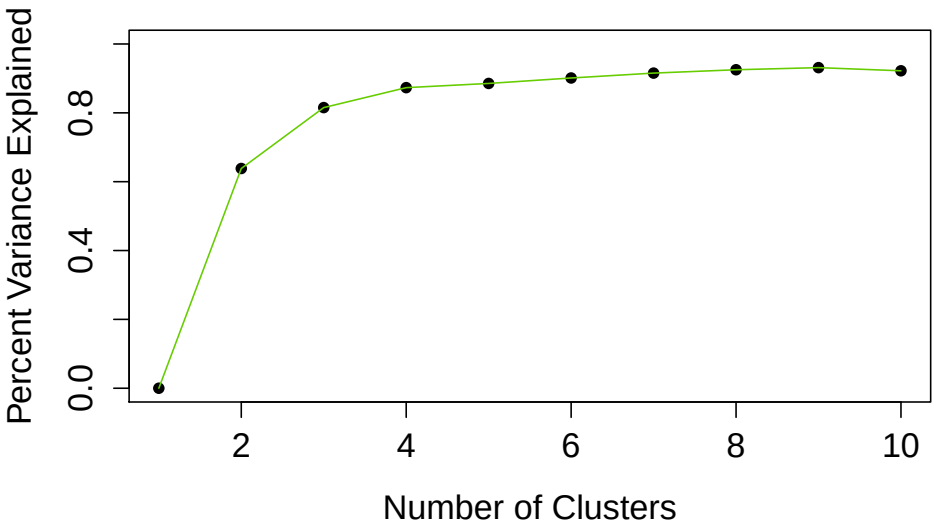

Fig. 5. Variance explained by using different numbers of types to summarize utility functions.

(3) *most cooperative*: those who cooperated in at least seven training observations (N=161).

We then performed several "one vs. all" classification tasks using logistic ridge regression with all of the survey and demographic items (plus squared terms and all two-way interactions) as features.[8] Training a model to predict whether or not a given participant was "least cooperative" (that is, giving a label of 1 when the participant was a "least cooperative" type and 0 otherwise) yielded an AUC of 0.54; predicting "intermediately cooperative" yielded an AUC of 0.54; and predicting "most cooperative" yielded an AUC of 0.53. Thus, the wide range of survey and demographic factors we considered provided almost no ability to predict cooperative type.[9] This result suggests that cooperativeness is a "natural kind," rather than being derivative of other personality features.

4. CONCLUSION

Here we have used machine learning to investigate the predictive power of social preference models when describing cooperative decision-making. Our results show that outcome-based social preference models can successfully predict cooperative behavior with high accuracy. This indicates that the game payoffs do have a considerable influence on game play: cooperation decisions are not largely random, nor driven by transient sentiments (e.g. mood). Instead, they reflect stable individual preferences over outcomes.

[8]The participants were randomly split into a training set (N=710) and test set (N=36), and this same test-train split was used for all analyses.

[9]Critically, we are performing out-of-sample prediction, rather than just asking how various measures correlate with cooperation within-sample as is typically done in social science research.

The fact that participant dummies alone have substantial predictive power highlights the importance of individual differences in disposition to cooperate. This finding is consistent with prior work finding strong correlations between an individual's play in different cooperation games [Balliet et al. 2009; Peysakhovich et al. 2014; Van Lange 1999; Yamagishi et al. 2013] and supports the notion of a "cooperative phenotype" introduced by [Peysakhovich et al. 2014].

However, the fact that including outcome-based features further increase predictive power above and beyond participant dummies shows that the game payoffs also matter. Individuals do not just differ in their general propensity to cooperate – rather, they differ in how they condition on outcomes. This combination of strong predictive power without conditioning on outcomes, but improved prediction when considering outcomes, suggests that simple decision heuristics which are insensitive to details of the situation [Bear and Rand 2016; Capraro et al. 2014; Rand 2016; Rand et al. 2014] are important for cooperation, but that deliberative sensitivity to payoff details (be it via deliberation or more sensitive heuristics) also plays an important role.

Furthermore, we show that this heterogeneity in social preferences can be well captured by just three distinct types: the generous who usually give, the selfish who rarely give, and the hyper-selfish who never give at all (i.e. homo economicus). Thus, while the representative agent assumption is not suitable for considering social preferences, it is also not necessary to go to the other extreme of specifying a unique set of preferences for every individual. At least within the current dataset, three types are sufficient to well characterize participant choices.

These findings add to a body of literature on the classification of social preferences, including work on "conditional cooperation" that examines how people respond to the behavior of others [Fischbacher et al. 2001], in contrast to our examination of unconditional play in 1-shot games; and work on "Social Value Orientation" where participants make unilateral choices between options that maximize the decision-maker's absolute payoff, the decision-maker's relative payoff, or the partner's payoff, and are classified accordingly as pro-self, competitive, or pro-social [Balliet et al. 2009; Van Lange 1999].

Finally, we find initial evidence that this variation in cooperativeness we observe represents a natural kind, rather than being derivative of other traits. We observe that the combination of a variety of survey measures related to personality and morality, as well as a wide range of demographics, do quite poorly at predicting participants' cooperative type (although it is of course possible that a different set of survey measures might have been more successful). This observation adds further weight to the usefulness of the cooperative phenotype as a fundamental feature of personality which is useful for explaining human prosociality.

In addition to generating insight regarding social preferences and cooperation, our results add to the accumulating body of support for the validity of small-stakes MTurk experiments.[10] In particular, there are two common concerns raised regarding economic game studies run on MTurk because the decisions involve very small amounts of money. The first concern is that because the decisions are not very monetarily consequential, participants will not take the experiments seriously and their behavior will be more noisy/random. The fact that we can predict out-of-sample play with high accuracy, however, shows that this is clearly not the case. The second concern is that because the stakes are low, it makes it "easier" to be prosocial (or conform to social expectations), such that cooperativeness will be over-estimated. Inconsistent with this

[10]Prior work has, for example, found very similar results when the same experiment was run on MTurk and in the physical lab with higher stakes [Horton et al. 2011; Peysakhovich and Karmarkar 2015; Rand et al. 2012; Suri and Watts 2011], and shown that complicated learning paradigms work on MTurk [Fudenberg and Peysakhovich 2014].

suggestion, however, we find extremely low giving rates: only 17.8% of the decisions in our dataset were cooperative. Thus our results suggest that MTurk participants do take their low-stakes decisions seriously.

While our results provide clear evidence for the power of heterogeneous outcome-based social preferences for predicting behavior, there are several caveats that merit further study. Most importantly, our design constrained participants to unilateral monetary interactions with anonymous strangers. In multilateral settings, or settings where information about the recipient's past behavior is available, we would expect non-outcome-based preferences (such as reciprocity [Levine 1998] and conditional cooperation [Fischbacher et al. 2001]) to play an important role in choice, as well as strategic reasoning. Although the inputs to these other social preferences are much harder to observe (and therefore to input as features), investigating how to harness machine learning to explore them is an important direction for future work (as has been done for strategic reasoning [Balliet et al. 2009; Van Lange 1999; Wright and Leyton-Brown 2012, 2014]). So too is investigating the power of outcome-based preferences, and the importance of individual differences, when predicting cooperation in naturalistic settings outside the lab [Kraft-Todd et al. 2015]. We hope that the approach we have introduced here will provide a framework for future investigations of the nature of human prosociality.

ACKNOWLEDGMENTS

Funding from the John Templeton Foundation and the VIA Institute is gratefully acknowledged.

REFERENCES

J. Andreoni and J. Miller. 2002. Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica* 70 (2002), 737–753.

D. Balliet, C. Parks, and J Joireman. 2009. Social value orientation and cooperation in social dilemmas: A meta-analysis. *Group Processes & Intergroup Relations* 12, 4 (2009), 533–547.

A. Bear and D. G. Rand. 2016. Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences* 70 (2016), 737–753.

G. S. Becker. 1976. *The economic approach to human behavior*. University of Chicago Press.

M. Blanco, D. Engelmann, and H. T. Normann. 2011. A within-subject analysis of other-regarding preferences. *Games and Economic Behavior,* 70, 2 (2011), 321–338.

G. E. Bolton and A. Ockenfels. 2000. ERC: A Theory of Equity, Reciprocity, and Competition. *The American Economic Review* 90, 1 (2000). http://www.jstor.org/stable/117286.

A. P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition* 30, 7 (1997), 1145–1159.

V. Capraro, J. J. Jordan, and D. G. Rand. 2014. Heuristics guide the implementation of social preferences in one-shot Prisoner's Dilemma experiments. *Scientific Reports* 4 (2014), 6790.

G. Charness and M. Rabin. 2002. Understanding Social Preferences with Simple Tests. *Quarterly Journal of Economics* 117, 3 (2002), 817–869.

S. Epstein, R. Pacini, V. Denes-Raj, and H. Heier. 1996. Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology* 71, 2 (1996), 390–405.

A. Fehr and K. M. Schmidt. 1999. A theory of fairness, competition and cooperation. *Quarterly Journal of Economics* 114, 3 (1999), 817–868.

E. Fehr, A. Klein, and K. M. Schmidt. 2004. Fairness and Incentives in a Multi–task Principal–Agent Model. *The Scandinavian Journal of Economics* 106, 3 (2004), 453–474.
E. Fehr, A. Klein, and K. M. Schmidt. 2007. Fairness and contract design. *Econometrica* 75, 1 (2007), 121–154.
U. Fischbacher, S. Gächter, and E. Fehr. 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters* 71, 3 (2001), 397–404.
R. Fisman, S. Kariv, and D. Markovits. 2007. Individual preferences for giving. *The American Economic Review,* 97, 5 (2007), 1858–1876.
J. Friedman, H. Trevor, and R. Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33, 1 (2010), 1–22. http://www.jstatsoft.org/v33/i01/
D. Fudenberg and E. S. Maskin. 1986. The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. *Econometrica* 54, 3 (1986), 533–554.
D. Fudenberg and A. Peysakhovich. 2014. Recency, Records and Recaps: Learning and Non-equilibrium Behavior in a Simple Decision Problem. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation (EC '14)*. ACM, New York, NY, USA, 971–986. DOI:http://dx.doi.org/10.1145/2600057.2602872
S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.
J. D. Greene, R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen. 2001. An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science* 293 (2001), 2105–2108.
A. E. Hoerl and R. W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 1 (1970), 55–67.
J. J. Horton, D. G. Rand, and R. J. Zeckhauser. 2011. The Online Laboratory: Conducting Experiments in a Real Labor Market. *Experimental Economics,* 14, 3 (2011), 399–425.
G. Kraft-Todd, E. Yoeli, S. Bhanot, and D. Rand. 2015. Promoting cooperation in the field. *Current Opinion in Behavioral Sciences* 3 (2015), 96–101.
D. K. Levine. 1998. Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics* 1, 3 (1998), 593–622.
M. A. Nowak and K. Sigmund. 1998. Evolution of indirect reciprocity. *Nature* 437, 7063 (1998), 1291–1298.
A. Persakhovich and D.G. Rand. 2015. Habits of Virtue: Creating Norms of Cooperation and Defection in the Laboratory. *Management Science* 62, 3 (2015), 631–647. DOI:http://dx.doi.org/10.1287/mnsc.2015.2168
C. Peterson and M. E. P. Seligman. *Character strengths and virtues: A classification and handbook*. American Psychological Association, Washington, DC.
A. Peysakhovich and U. Karmarkar. 2015. Asymmetric Effects of Favorable and Unfavorable Information on Decision Making Under Ambiguity. *Management Science* (2015). DOI:http://dx.doi.org/10.1287/mnsc.2015.2233,
A. Peysakhovich and J. Naecker. 2015. Using Methods from Machine Learning to Evaluate Models of Human Choice. (2015).
A. Peysakhovich, M. A. Nowak, and D. G. Rand. 2014. Humans Display a 'Cooperative Phenotype' that is Domain General and Temporally Stable. *Nature Communications* 5 (2014), 4939.
D.G. Rand. 2016. Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science,* (2016). In press.
D. G. Rand, S. Arbesman, and N. A. Christakis. 2011. Dynamic social networks pro-

mote cooperation in experiments with humans. 108, 48 (2011), 19193–19198.
D. G. Rand, J. D. Greene, and M. A Nowak. 2012. Spontaneous Giving and Calculated Greed. *Nature* 489, 7416 (2012), 427–430.
D. G. Rand and G. T. Kraft-Todd. 2014. Reflection Does Not Undermine Self-Interested Prosociality. *Frontiers in Behavioral Neuroscience,* 8, 300 (2014).
D. G. Rand and M. A. Nowak. 2013. Human Cooperation. *Trends in cognitive sciences* 17, 8 (2013), 413–425.
D. G. Rand, A. Peysakhovich, G. T. Kraft-Todd, G. E. Newman, O. Wurzbacher, M. A. Nowak, and J. D. Greene. 2014. Social Heuristics Shape Intuitive Cooperation. *Nature Communications* 5, 3677 (2014).
S. Suri and D. J. Watts. 2011. Cooperation and Contagion in Web-Based, Networked Public Goods Experiments. *PLoS ONE* 6, 3 (2011). DOI:http://dx.doi.org/10.1371/journal.pone.0016836
P. A. M. Van Lange. 1999. The Pursuit of Joint Outcomes and Equality in Outcomes: An Integrative Model of Social Value Orientation. *Journal of Personality and Social Psychology* 77, 2 (1999), 337–349. http://www.jstor.org/stable/23723482
J. R. Wright and K. Leyton-Brown. 2012. Behavioral game theoretic models: a Bayesian framework for parameter analysis. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. International Foundation for Autonomous Agents and Multiagent Systems, 921–930.
J. R. Wright and K. Leyton-Brown. 2014. Level-0 meta-models for predicting human behavior in games. In *Proceedings of the fifteenth ACM conference on Economics and computation*. ACM, 857–874.
T. Yamagishi, N. Mifune, Y. Li, M. Shinada, H. Hashimoto, Y. Horita, A. Miura, K. Inukai, S. Tanida, T. Kiyonari, H. Takagishi, and D. Simunovic. 2013. Is behavioral pro-sociality game-specific? Pro-social preference and expectations of pro-sociality. *Organizational Behavior and Human Decision Processes* 120, 2 (2013), 260–271.