Essays on Belief Formation and Pro-Sociality

EFI THE ECONOMIC RESEARCH INSTITUTE

EFI Mission

EFI, the Economic Research Institute at the Stockholm School of Economics, is a scientific institution that works independently of economic, political and sectional interests. It conducts theoretical and empirical research in the management and economic sciences, including selected related disciplines. The Institute encourages and assists in the publication and distribution of its research findings and is also involved in the doctoral education at the Stockholm School of Economics. At EFI, the researchers select their projects based on the need for theoretical or practical development of a research domain, on their methodological interests, and on the generality of a problem.

Research Organization

The research activities at the Institute are organized into 20 Research Centres. Centre Directors are professors at the Stockholm School of Economics.

EFI Research Centre:	*Centre Director:*
Management and Organization (A)	Sven-Erik Sjöstrand
Centre for Entrepreneurship and Business Creation (E)	Carin Holmquist
Public Management (F)	Nils Brunsson
Information Management (I)	Mats Lundeberg
Centre for People and Organization (PMO)	Andreas Werr
Centre for Innovation and Operations Management (T)	Pär Åhlström
Centre for Media and Economic Psychology (P)	Richard Wahlund
Centre for Consumer Marketing (CCM)	Magnus Söderlund
Centre for Information and Communication Research (CIC)	Per Andersson
Marketing, Distribution and Industry Dynamics (D)	Björn Axelsson
Centre for Strategy and Competitiveness (CSC)	Örjan Sölvell
Accounting and Managerial Finance (B)	Johnny Lind
Centre for Financial Analysis and Managerial Economics in Accounting (BFAC)	Kenth Skogsvik
Finance (FI)	Clas Bergström
Centre for Health Economics (CHE)	Magnus Johannesson
International Economics and Geography (IEG)	Mats Lundahl
Economics (S)	Paul Segerstrom
Economic Statistics (ES)	Jan Eklöf
Centre for Business Law (RV)	Erik Nerep
Centre for Tax Law (SR)	Bertil Wiman

Chair of the Board: Professor Carin Holmquist
Director: Associate Professor Filip Wijkström

Address

EFI, Box 6501, SE-113 83 Stockholm, Sweden • Website: www.hhs.se/efi/
Telephone: +46(0)8-736 90 00 • Fax: +46(0)8-31 62 70 • E-mail: efi@hhs.se

Essays on Belief Formation and Pro-Sociality

Erik Mohlin

EFI THE ECONOMIC RESEARCH INSTITUTE

Dissertation for the Degree of Doctor of Philosophy, Ph.D.
Stockholm School of Economics

KEYWORDS: Categorization; Communication; Dictator game; Evolution; Learning; Level-k; Prediction; Priors; Social Norms; Theory of Mind; Ultimatum game.

ISBN 978-91-7258-825-7

PRINTED BY:
Intellecta Infolog, Göteborg 2010

DISTRIBUTED BY:
EFI, The Economic Research Institute
Stockholm School of Economics
P O Box 6501, SE-113 83 Stockholm
www.hhs.se/efi

Företal

Föreliggande arbete utgör resultatet av ett forskningsprojekt som bedrivits vid Ekonomiska forskningsinstitutet vid Handelshögskolan i Stockholm. Som brukligt är vid Ekonomiska forskningsinstitutet har författaren haft full frihet att självständigt utforma projekt- och resultatredovisning. Institutet är tacksamt för det finansiella stöd som möjliggjort projektets genomförande.

Stockholm

Filip Wijkström	Paul Segerstrom
Docent och chef för	Professor och prefekt för
Ekonomiska forskningsinstitutet	Nationalekonomiska institutionen

To my parents, Bodil and Bengt

Contents

Introduction

This thesis consists of four independent papers, ordered chronologically with respect to when they were initiated. The first two papers use experimental methods to study pro-social behaviors. The other two use theoretical methods to investigate questions about belief formation. The first paper investigates the effect on communication on altruism. The second paper is about the neurological basis for the tendency to punish norm violators. The third paper explores how evolution might shape the ways in which we form expectations about other people's behavior in strategic interactions. The fourth paper is concerned with how categorical thinking is used to form beliefs.

1. Communication and Generosity

It is well-established that pre-play communication increases the degree of generosity or cooperation in many kinds of experimental games (Sally 1995, Ledyard 1995, Crawford 1998, and Camerer 2003). One explanation for this fact is that communication enables people to *coordinate* their behavior. For instance, if experimental subjects play a social dilemma game they might prefer to cooperate if other people cooperate but prefer not to cooperate if other people fail to cooperate (Sen 1967). In the absence of communication people might not dare to cooperate because they fear that other people will not cooperate. Communication may enable to reassure people that everyone else will cooperate, thereby making it optimal for everyone to cooperate.

In order to test whether coordination can fully explain the positive effect of communication on pro-social behaviors, Magnus Johannesson and I investigated the effect of different kinds of communication in a dictator game. In the dictator game only one person gets to act so there is no scope for coordination. The results are presented in our paper "Communication: Content or Relationship?"

In the basic experiment (the control), subjects in one room are *dictators* and subjects in another room are *recipients*. The subjects are anonymous to each other throughout the whole experiment. Each dictator gets to allocate a sum of 100 SEK between herself and an unknown recipient in the other room. The dictators can keep any amount they want, but some still chose to donate some of their 100 SEK.

In the first treatment we allow each recipient to send a free-form message to his dictator counterpart, before the dictator makes her allocation decision (still preserving anonymity). We find that this significantly increases donations. This demonstrates that the effect of communication on generosity is not exclusively due to coordination.

In order to separate the effect of the *content* of the communication, from the *relationship*-building effect of communication, we carry out a third treatment, where we take the messages from the previous treatment and give each of them to a dictator in this new treatment. The dictators are informed that the recipients who wrote the messages are not the recipients they will have the opportunity to send money to. We find that this still increases donation compared to the baseline but not as much as in the other treatment. This suggests that both the impersonal content of the communication and the relationship effect matters for donations.

One can speculate about the mechanism underlying the observed behavior. The content of the communication may affect the fairness norm or the cost of deviating from this norm (in line with Rabin (1994) or Konow (2000). There may also be a relationship specific effect of communication; communication may for instance increase the empathy for the recipient or decrease the social distance between the dictator and the recipient (Bohnet and Frey 1999, Hoffman et al. 1996).

2. The Neurological Basis of Punishment of Norm Violations

Social norms are fundamental for regulating human relationships and organizing human societies (Elster 1989, Ostrom 2000). Norms that govern cooperation are of particular importance; humans cooperate with genetically unrelated individuals to an extent that is unparalleled in the animal world (Hammerstein 2003). It seems that an important mechanism for maintaining cooperation and compliance with social norms is the human propensity to punish those who violate norms of cooperation (Yamagishi 1986, Fehr and Gächter 2000). Cooperative behavior is rewarded with cooperation, and a failure to cooperate is punished.

The Ultimatum Game (Güth et al. 1982) is a suitable experimental setting for studying this mechanism. In the Ultimatum Game, a *proposer* proposes a way to divide a fixed sum of money. The *responder* accepts or rejects the proposal. If the proposal is accepted the proposed split is realized and if the proposal is rejected both subjects gets zero. The payoff-maximizing strategy of the responder is to accept all offers no matter how small. Given this behavior on part of the responder, the optimal strategy for the proposer is to offer almost nothing. However, most people do not behave in this way. Unfair offers are frequently rejected, and this finding is robust with respect to learning effects, stake size, and other manipulations (Camerer 2003). Cultural variation has been documented but the general propensity to punish norm violators seems to be universal, and has a substantial genetic component (Heinrich et al. 2005, Wallace et al. 2007).

In the context of repeated interactions, costly punishment can be rational from a selfish point of view – and as fitness maximizing from an evolutionary perspective – since the immediate cost of punishment can be balanced by the long term benefit of maintaining good reciprocal relationships and gaining a beneficial reputation (Luce and Raiffa 1957, Trivers 1971, Kreps et al. 1982, and Sigmund and Nowak 1998). However, in the Ultimatum Game players only meet once so there is no scope for reciprocity or reputation building.[1] Instead the decision to reject must be driven by some direct

[1] It seems that this behavior can not be evolutionarily adaptive, unless there are high costs of discriminating between repeated and one-shot interactions. Though see Schaffer (1989), Huck and Oechssler (1999), and Binmore et al. (1995).

reward that is attached to the act of punishment, as evidenced by de Quervain et al. (2004) and as described by theories of reciprocal preferences (Rabin 1993, Charness and Rabin 2002). More generally, evolution may have endowed humans with automatic responses that are adapted to the context of repeated non-anonymous interactions. These automatic responses may be present in the Ultimatum Game even though the interaction is one-shot. Thus, a responder receiving an unfair offer faces a trade-off between two conflicting forces: On the one hand punishment implies a monetary loss. On the other hand, punishment is associated with a direct reward, possibly linked to hard-wired concerns about reciprocity and reputation.

Previous studies have found somewhat mixed evidence on the neural basis of the rejection behavior in the Ultimatum Game. Sanfey et al. (2003) found increased activation in the dorsolateral prefrontal cortex (DLPFC) as a response to unfair offers in the Ultimatum Game. They also found activation in the anterior insula, and in the anterior cingulated cortex (ACC). According to their interpretation, the insula activity represents an emotional impulse to reject unfair offers, whereas the DLPFC represents a cognitive process that controls the impulse to reject, and the activation of ACC is due to these conflicting motives. Another previous study (Knoch et al. 2006) comes to a different conclusion. They find that inhibiting the right DLPFC with transcranial magnetic stimulation (rTMS) lowers rejection rates. An interpretation consistent with this finding is that the DLPFC activity represents a cognitive process that exerts top-town control of an emotional impulse to accept all offers.

In the paper "Limbic justice – Amygdala Drives Rejection in the Ultimatum Game" (Co-authors Katarina Gospic, Peter Fransson, Predrag Petrovic, Magnus Johannesson, and Martin Ingvar) we present the results from an experiment on Ultimatum Game responder behavior, where the rejection decision is manipulated with the help of a pharmacological intervention. In our experiment thirty-five subjects were randomly allocated to receive either the benzodiazepine oxazepam or a placebo substance, and then played the Ultimatum Game in the responder role, while lying in and fMRI camera. We found that the rejection rate is significantly lower in the treatment group than in the control group. Moreover amygdala was relatively more activated in the placebo group than in the oxazepam group for unfair offers. This is mirrored by differences in activation in the medial prefrontal cortex (mPFC) and right ACC. Amygdala activity corresponds to an increased rejection rate also within the placebo group. Importantly no effects related to rejection were observed in dlPFC and anterior insula, seemingly in contrast with the previously mentioned studies.

We argue that the activation that Sanfey et al. (2003) detect in the anterior insula is merely a lingering trace of a process that we are able to record at its origin in the amygdala. The reason is that we use a more clearly defined onset time for measuring the reaction to unfair proposals. Furthermore we argue that the finding in Knoch et al. (2006) is likely to be an artefact of the technology they use. In conclusion, our findings suggest that the automatic and emotional response to unfairness, or norm violations, are driven by a phylogenetically old structure (amygdala) and that balancing of such automatic behavioral responses is associated with cortical structures that are phylogenetically younger (prefrontal cortex).

3. Models for Forming Beliefs about Other Peoples' Beliefs

In order to decide what strategy to choose, a player needs to form beliefs about what other players will do. This requires the player to have a model of how other people form beliefs – what psychologists call a theory of mind (Premack and Wodruff 1978). In the paper "Evolution of Theories of Mind" I study the evolution of players' models of how other players think.

A set of behaviors in a game is said to constitute a *Nash equilibrium* of the game if (a) all players do what they find best given their beliefs and (b) their beliefs are correct. Naturally, the second condition will generally not be satisfied unless the players have had opportunity to learn how other people behave. Hence when people play a game for the first time, their behavior is more successfully predicted by the *level-k* (Stahl and Wilson 1995, Nagel 1995) and cognitive hierarchy models (Camerer et al. 2004), or models of noisy introspection (Goeree and Holt 2004). According to these models, people think in a limited number of steps, when they form beliefs about other peoples' behavior. Moreover, people differ with respect to how they form beliefs. The heterogeneity is represented by a set of cognitive types $\{0, 1, 2, ...\}$, such that higher types form more sophisticated beliefs. Type 0 does not form any beliefs and randomizes uniformly over the strategy space. According to the level-k model, an individual of type k believes that everyone else belongs to type $k-1$. All types k best respond given their beliefs, and have identical preferences. Empirically one finds that most experimental subjects behave as if they are of type 1 or 2, and individuals of type 3 and above are very rare (Costa-Gomes and Crawford 2006, Camerer 2003).

When people have had some experience with playing a particular game a number of times against a group of people, they may use their experience to predict what the opponents will do in the future. In order to make such predictions a player needs to have some model of how the other players use their experience to form beliefs. That is, the player needs a model of how other players think. One prominent model of learning is *fictitious play* (see Fudenberg and Levine 1998). This model postulates that all individuals believe that the future will be like the past, and best respond to the average of past play. However, if some individuals follow this rule, it is natural to hypothesize that some more sophisticated individuals could understand that other individuals behave in accordance with fictitious play. These more sophisticated individuals would then play a best response to the best response to the average of past play. And it seems quite possible that some individuals think yet another step and play a twice iterated best response to the average of past play. Continuing in this way we arrive at a hierarchy of types $\{1, 2, ...\}$, using increasingly complex models of how other people learn. I refer to the resulting model as *heterogeneous fictitious play.* A related model is proposed by Stahl (1999), and subjected to experimental testing in (Stahl 2000).

The level-k and cognitive hierarchy models, as well as fictitious play, implicitly assume that players lack specific information about the cognitive types of their opponents. I extend these models to allow for the possibility that types are partially observed. Such an extension is essential in order to capture situations where an unfamiliar game is played by individuals who have some information about their opponents' ways of forming beliefs.

I study evolution of types in a number of games separately. In contrast to most of the literature on evolution and learning, I also study the evolution of types across different games. I show that an evolutionary process, based on payoffs earned in different games, both with and without partial observability, can lead to a polymorphic population where relatively unsophisticated types survive, often resulting in initial non-Nash behavior.

Two important mechanisms behind these results are the following: (i) There are games, such as the Hawk-Dove game, where there is an advantage of not thinking and behaving like others, since choosing the same action as the opponent yields an inefficient outcome. This mechanism is at work even if types are not observed. (ii) If types are partially observed then there are Social dilemmas where lower types may have a commitment advantage; lower types may be able to commit to strategies that result in more efficient payoffs.

From an evolutionary perspective, the potential advantage of a better theory of mind has to be weighted against the cost of increased reasoning capacity. I abstract from such costs in the formal analysis, but note that they limit the survival chances of higher types.

According to the prominent "social brain", or "Machiavellian intelligence", hypothesis (e.g. Humphrey 1976, Alexander 1990, Byrne and Whiten 1998), the extraordinary cognitive abilities of humans, evolved as a result of the demands of social interactions, rather than the demands of the natural environment. In a single person decision problem there is a fixed benefit of being smart, but in a strategic situation it may be important to be smarter than the opponent. The results in this paper complement the social brain hypothesis by suggesting mechanisms that may sustain heterogeneity with respect to theory of mind abilities.

4. Beliefs Formed on the Basis of Categorizations

The importance of categorical reasoning in human cognition is well-established in psychology and cognitive science (Laurence and Margolis 1999, Murphy 2002). One of the most important functions of categorization is to facilitate prediction (Anderson 1990).

Prediction on the basis of categorical reasoning is relevant when one has to predict the value of a variable on the basis of one's previous experience with similar situations, but where the past experience does not include any situation that was identical to the present situation in all relevant aspects. In such situations one can classify the situation as belonging to some category, and use the past experiences in that category to make a prediction about the current situation.

In the paper “Optimal Categorization” I provide a model of categorizations that are optimal in the sense that they minimize prediction error. Both costs and benefits are derived endogenously from the objective of making accurate predictions. The advantage of fine grained categorizations is that objects in a category are similar to each other. The advantage of coarse categorizations is that a prediction about a category is based on a large number of observations.

Why should we be interested in optimal categorizations? From an evolutionary perspective we would expect humans to have developed categories that generate predictions which induce behavior that maximize fitness. It seems reasonable to assume that fitness is generally increasing in how accurate the predictions are. For instance, a subject encountering a poisonous plant will presumably be better off if she predicts that the plant is indeed poisonous, rather than nutritious. For this reason we would expect to find that humans employ categorizations that are at least approximately optimal, in the sense that they minimize prediction error.

As an illustration, think of color concepts. The subset of the spectrum of electromagnetic radiation that is visible to the human eye allows for infinitely fine grained distinctions. However, in every day reasoning and discourse we seem to employ only a coarse color classification, using words such as red, green, turquoise, etcetera. We could have sliced up the space of colors differently. Presumably the color categorizations that were developed and passed on to new generations were successful in the kind of environments that we faced.

In the model a subject starts out with a categorization that she has learnt or inherited early in life. The categorization divides the space of objects into categories. In the beginning of each period, the subject observes a two-dimensional object in one dimension, and wants to predict the object's value in the other dimension. She has a data base of objects that were observed in both dimensions in the past. The subject determines what category the new object belongs to on the basis of observation of its first dimension. She predicts that its value in the second dimension will be equal to the average value among the past observations in the corresponding category. At the end of each period the second dimension is observed, and the observation is stored in the data base.

The main result is that the optimal number of categories is determined by a trade-off between (a) decreasing the size of categories in order to enhance category homogeneity, and (b) increasing the size of categories in order to enhance category sample size. In other words, the advantage of fine grained categorizations is that objects in a category are similar to each other. The advantage of coarse categorizations is that a prediction about a category is based on a large number of observations, thereby reducing the risk of over-fitting. Comparative statics reveal how the optimal categorization depends on the number of observations as well as on the frequency of objects with different properties. The set-up does not presume the existence of an objectively true categorization "out there". The optimal categorization is a framework we impose on our environment in order to predict it.

The dominant view within psychology is that the number of categories (the coarseness of the categorization) is determined by another trade-off (Medin 1983). Like in my paper, the benefit of small categories is supposed to be within-category homogeneity of objects. But, unlike this paper, the benefit of having a fewer larger categories is supposed to be that one needs to observe fewer properties of an object in order to categorize it as belonging to a large category. A virtue of the explanation put forward in this paper is that it connects a main purpose of categorization, namely prediction, both with the value of many small categories and with the value of a few large categories.

References

Alexander, R. D. (1990), 'How did Humans Evolve? Reflections on the Uniquely Unique Species', *University of Michigan Museum of Zoology Special Publication* No 1.

Anderson, J. R. (1990), *The Adaptive Character of Thought*, Erlbaum, Hillsdale, NJ.

Binmore, K, Gale, J. and Samuelson, L. (1995), 'Learning to be imperfect: the Ultimatum Game', *Games and Economic Behavior*, 8, 56-90.

Bohnet, I., and Frey, B.S., (1999), 'The sound of silence in prisoner's dilemma and dictator games', *Journal of Economic Behavior and Organization* 38, 43-57.

Byrne, R. W. and Whiten, A. (1998), *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*, Oxford University Press, Oxford.

Camerer, C. (2003), *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton University Press, Princeton, NJ.

Camerer, C. F., Ho, T.-H. and Chong, J.-K. (2004), 'A Cognitive Hierarchy Model of Games', *Quarterly Journal of Economics* 199, 861-898.

Charness G, & Rabin M (2002), 'Understanding social preferences with simple tests', *Quarterly Journal of Economics* 117, pp. 817-869.

Costa-Gomes, M. A. and Crawford, V. P. (2006), 'Cognition and Behavior in Two-Person Guessing Games: an Experimental Study', *American Economic Review* 96, 1737-1768.

Crawford, V., (1998), 'A survey of experiments on communication via cheap talk', *Journal of Economic Theory* 78, 286-298.

Elster, J. (1989), *The Cement of Society – A Study of Social Order*, Cambridge U P, Cambridge.

Fehr, E. and Gächter, S. (2000), 'Cooperation and punishment in public goods experiments', *American Economic Review* 90, 980–994.

Fehr, E. and Gächter, S. (2002), 'Altruistic Punishment in Humans', *Nature* 415, 137-140

Fudenberg, D. and Levine, D. K. (1998), *The Theory of Learning in Games*, MIT Press, Cambridge, MA.

Goeree, J. K. and Holt, C. A. (2004), 'A model of noisy introspection', *Games and Economic Behavior* 46(2), 365-382.

Güth, W., Schmittberger. R., and Schwarze, B. (1982), 'An experimental analysis of ultimatum bargaining', *Journal of Economic Behavior and Organization* 3: 367–388.

Hammerstein, P. (ed.) (2003), *Genetic and Cultural Evolution of Cooperation.* Cambridge: The MIT Press.

Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, et al. (2005), '"Economic man" in crosscultural perspective: behavioral experiments in 15 small-scale societies' *Behavioral and Brain Science* 28, 795–815.

Hoffman, E., McCabe, K., Smith, V.L., (1996), 'Social distance and other regarding behavior in dictator games', *American Economic Review* 86, 653 -660.

Huck, S. and Oechssler, J. (1999), 'The indirect evolutionary approach to explaining fair allocations', *Games and Economic Behavior* 28, 13-24.

Humphrey, N. K. (1976), 'The social function of intellect', in P. P. G. Bateson and R. A. Hinde, eds, *Growing Points in Ethology*, Cambridge University Press, Cambridge, 303-317.

Knoch D, Pascual-Leone A, Meyer K, Treyer V, Fehr E. (2006), 'Diminishing reciprocal fairness by disrupting the right prefrontal cortex', *Science* 314, 829-832.

Konow, J., (2000), 'Fair shares: accountability and cognitive dissonance in allocation decisions', *American Economic Review* 90, 1072-1091.

Kreps, D., P. Milgrom, J. Roberts, and R. Wilson. (1982), 'Rational Cooperation in the Finitely Repeated Prisoners' Dilemma', *Journal of Economic Theory* 27, 245-52.

Laurence, S. and Margolis, E. (1999), 'Concepts and Cognitive Science', in E. Margolis and S. Laurence, eds, *.Concepts: Core Readings*, MIT Press, Cambridge, MA, pp. 3-81.

Ledyard, J.O., (1995), 'Public goods: a survey of experimental research', In: Kagel, J.H., Roth, A.E. (Eds.). *Handbook of Experimental Economics.* Princeton: Princeton University Press, 111-194.

Luce, D och H Raiffa (1957), *Games and Decisions*, Wiley, New York.

Medin, D. L. (1983), 'Structural Principles of Categorization', in B. Shepp and T. Tighe, eds, *Interaction: Perception, Development and Cognition*, Erlbaum, Hillsdale, NJ, 203-230.

Murphy, G. L. (2002), *The Big Book of Concepts*, MIT Press, Cambridge, MA.

Nagel, R. (1995), 'Unraveling in Guessing Games: An Experimental Study', *American Economic Review* 85, 1313-1326.

Nowak MA, K Sigmund (1998), 'Evolution of indirect reciprocity by image scoring' *Nature* 393, 573-577.

Ostrom E (2000), "Collective action and the evolution of social norms", *Journal of Economic Perspectives* 14(3), 137-158..

Premack, D. and Wodruff, G. (1979), 'Does the Chimpanzee have a Theory of Mind', *Behavioral and Brain Sciences* 1, 515-526.

Rabin, M. (1993), 'Incorporating Fairness into Game Theory and Economics', *American Economic Review* 83(5), 1281-1302.

Rabin, M., (1994), 'Cognitive dissonance and social change', *Journal of Economic Behavior and Organization* 23, 177-194.

Sally, D., (1995), 'Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958-1992', *Rationality and Society* 7, 58-92.

Sen, A.K., (1967), 'Isolation, assurance and the social rate of discount', *Quarterly Journal of Economics* 81, 112-124.

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003), 'The neural basis of economic decision-making in the ultimatum game', *Science* 300, 1755-1758.

Schaffer, M. E. (1988), 'Evolutionarily stable strategies for finite population and a variable contest size', *Journal of Theoretical Biology* 132, 469-478.

Stahl, D. O. (1993), 'Evolution of Smart$_n$ Players', *Games and Economic Behavior* 5, 604-617.

Stahl, D. O. (1999), 'Evidence based rules and learning in symmetric normal-form games', *International Journal of Game Theory* 28(1), 111-130.

Stahl, D. O. (2000), 'Rule Learning in Symmetric Normal-Form Games: Theory and Evidence', *Games and Economic Behavior* 32, 105-138.

Stahl, D. O. and Wilson, P. W. (1995), 'On Players Models of Other Players: Theory and Experimental Evidence', *Games and Economic Behavior* 10, 218-254.

de Quervain DJF, Fischbacher U, Treyer V, Schellhammer M, Schnyder U, Buck A, Fehr E. (2004), 'The neural basis of altruistic punishment' *Science* 305, pp. 1254-1258.

Trivers, R. (1971), 'The evolution of reciprocal altruism', *Quarterly Review of Biology* 46, 35–56.

Wallace, B., Cesarini, D., Lichtenstein, P. and Johannesson, M. (2007), 'Heritability of Ultimatum Game Responder Behavior', *Proceedings of the National Academy of Sciences* 104, 15631-15634.

Yamagishi, T. (1986), 'The provision of a sanctioning system as a public good', *Journal of Personality and Social Psychology* 51, 110–116.

Acknowledgement

This thesis was made possible with the advice and support from many people.

Magnus Johannesson has been my main supervisor and my co-author on two of the papers in this thesis. He has always been very generous with his time and advice. His advice builds on extensive knowledge of experimental economics and an impeccable judgment regarding experimental methodology. In addition to this Magnus is a very kind and considerate person.

Jörgen Weibull became my extra supervisor at the end of my second year. During numerous sessions he has guided me in the art of crafting economic models. Without the backing of his mathematical knowledge and his theoretical precision I would have been lost at many points. Jörgen was also very encouraging when I doubted that some of my projects were worth pursuing.

Although not a supervisor, Tore Ellingsen has also been an important figure. Talking to him is always extremely inspiring. He has also been very supportive of some of my "non-economic" ideas.

Karl Wärneryd advised me on my first attempts to do economic theory and apply it to evolutionary questions. He also encouraged me to finish and publish my project.

I have had many stimulating discussion with Ola Andersson about game theory and behavioral economics, and he has provided very detailed and valuable feedback on my theoretical papers. He has also put up with me constantly dropping by his office to discuss research problems and to exchange gossip.

Other members of faculty that made my life at the department interesting and enjoyable include David Domeij, Martin Flodén, Rickard Friberg, Juanna Joensen, and Mark Voornefeld. Thanks also to Stefano Demichelis, a regular visitor at the department, and Per Hedberg, at the department of the department of marketing and strategy.

Many thanks go to the efficient and friendly administrative staff, consisting of Anna Angerman, Carin Blankswärd, Ritva Kiviharju, Anneli Sandblad, and Lilian Öberg. The department would not function a day without them.

I was fortunate to have the opportunity to make two academic visits to Boston. This spring I spent some very stimulating months at Harvard University. I wish to thank

Drew Fudenberg who hosted me and gave me valuable comments on my research. At an earlier stage I spent two months at Boston University. Bart Lipman kindly hosted me there.

I thank my co-authors at Karolinska Institutet; Peter Fransson, Katarina Gospic, Martin Ingvar, and Predrag Petrovic.

Financial support from the Jan Wallander and Tom Hedelius Foundation is gratefully acknowledged.

Many fellow PhD students have inspired me in the process of writing this thesis. Moreover they managed to turn the, sometimes quite rough, journey into a lot of fun.

Innumerable discussions with Robert Östling have been extremely important in shaping my views on economics, game theory and behavioral economics. He has also been a very good friend. We have not yet managed to write a joint paper, but hopefully that day will come.

Eva Ranehill and I started the PhD program at the same time and she has been my roommate for the last years. I am very grateful for having had her as a friend to share the ups and downs of PhD life with. It really made a difference.

Per Sonnerby was instrumental in my decision to study economics at all, and provided a role model for how to combine economic rationality with social consciousness. Andreas Müller was my room mate and study partner during the first tough year. Emma von Essen has been an occasional but always very welcome room mate. Erik Lindqvist was very kind and considerate, and even let me inherit his coffee mug. Linus Siming made me laugh, and was always willing to share his classified informed about economists. André Romahn also made me laugh and, among other things, came up with the great idea of moving the economics department to Nordiska museet. Anna Dreber generously shared both her network and her knowledge of biology and economics. Björn Wallace was always willing to provoke, and enlighten me, on matters as diverse as evolutionary theory and politics. Mark Bernhard and Margherita Bottero provided me with many interesting discussion about behavioral economics, economic theory, and philosophy.

Many other PhD students deserve to be mentioned for contributing to the academic and social environment at the department: Johan Almenberg, Axel Bernergård, Göhkan Buturak, Max Elger, Palle Elger, Ronny Freier, Sara Formai, Karin Hederos Eriksson, Amanda Jakobsson, Karen Khachatryan, Tobias Laun, Henrik Lundvall (see figure 1 in chapter 2), Kristin Magnusson, Elena Mattana, Damian Migueles Chazarreta, Anna Sandberg, Ignat Stepanok, Björn Tyrefors, Alberto Vesperoni, and Nick Vikander.

My personal friends (you know who you are) have been an invaluable source of encouragement, common sense, endless discussions, and laughs.

I wish to express my profound gratitude towards my family. My mother and father, for their constant support and for always having believed in me. My sister Charlotta, my staunchest supporter, who always reminds me to pursue the kind of research that fascinates me, and not to follow the mainstream.

Finally, my deepest thank goes to Nånnis. Without her patient support and love, I would not have been able to accomplish this. This publication should really be referred to as Brauner & Mohlin (2010).

PAPER 1

Communication: Content or Relationship?

Erik Mohlin, Magnus Johannesson

Abstract. We investigate the effect of anonymous communication on generosity in a dictator game. One-way written communication from the recipient is compared with no communication. Communication increases donations by more than 70 percent ($p<0.05$). To separate the effect of the content of the communication from the "relationship effect" of communication, a third treatment is carried out with one-way communication from third-parties (as messages from the recipients in the second treatment). In this third treatment, the donations are about 40 percent higher than in the treatment with no communication ($p<0.10$), suggesting that the impersonal content of the communication affects donations.

1. Introduction

Pre-play communication has been found to increase cooperative and other-regarding behavior in various experimental games (see for instance the overviews of experimental results in Ledyard (1995), Crawford (1998), and Camerer (2003)). In a meta-analysis of prisoners' dilemma experiments, Sally (1995) found that communication was the single most effective cooperation-increasing factor in one-shot as well as in repeated games. The effect of communication is not restricted to prisoners' dilemmas and related games, but is also found in different bargaining games. For example, Frey and Bohnet (1995) found that face-to face communication leads to more generous offers in dictator and ultimatum games, and Valley et al. (2002) found that written anonymous communication and face-to-face communication increase trade in a double-auction.

This paper explores some hypotheses about why communication influences behavior. One possible explanation is reputation building. If the communication involves revelation of players' identities it might be rational for a selfish person to behave in an other-regarding manner in order to build a good reputation. However, reputation

The authors thank two anonymous referees and Tore Ellingsen for helpful comments and Jon Fahlander, Martin Gemzell, Erik Lindqvist, Joel Malmqvist, Kalle Nilsson, Ingvar Strid, Olof Sundblad, Niklas Zethraeus and Robert Östling for research assistance. We also thank The Jan Wallander and Tom Hedelius Foundation and the Swedish Research Council for financial support.

building cannot explain why anonymous communication affects behavior, as is evidenced in several studies (e.g. Frohlich and Oppenheimer 1998, Bochet et al. 2002, Brosig et al. 2004, Ellingsen and Johannesson 2004).

One plausible explanation for the effect of anonymous communication is coordination. If there are multiple equilibria in a game, what equilibrium will be played depends on the players' expectations about other players' choices. A possible effect of communication is to influence these expectations and thereby behavior. However, there is only one equilibrium in the prisoners' dilemma if players are selfish, and choosing the strictly dominant defection strategy is not dependent on the behavior or rationality of the other players.

To accommodate the fact that people do not always defect in anonymous one-shot prisoners' dilemmas, one can invoke some kind of social preferences (e.g. Rabin 1993, Fehr and Schmidt 1999, Bolton and Ockenfels 2000, Charness and Rabin 2002). Doing so can create multiple equilibria and a role for communication to influence behavior. For example, Sen (1967) shows that a prisoner's dilemma in material payoffs might be a stag hunt game in subjective payoffs when players have preferences for cooperating as long as everyone else cooperates. In the stag hunt game there are two equilibria, and there is room for coordination on the Pareto-optimal rather than the inefficient equilibrium. That communication improves coordination is consistent with theory and experimental evidence (Crawford 1998, Valley et al. 2002).

Coordination may be important in games with strategic interactions such as the prisoners' dilemma. However, it would also be interesting to test whether communication has an effect in a non-strategic environment such as the dictator game. In the dictator game, one person (the dictator) decides how to split a sum of money between him/herself and another person (the recipient). With the exception of a study by Charness and Rabin (2005) we have not found any studies of anonymous communication in dictator games. Charness and Rabin (2005) allowed recipients to communicate their preference between two possible allocations before the dictator chose between the allocations. They find that behavior depends significantly on the preference expressed by the recipient.[1]

We compare an ordinary dictator game with a treatment that allows the recipient to send a written message to the dictator. Both treatments are double-blind (Hoffman et al. 1994). We find that communication increases the average donation by more than 70% ($p<0.05$).

[1] In gift exchange games and trust games it has also been found that communication in the form of desired effort levels or back transfers can affect behavior (Fehr et al. 2001, Fehr and Gächter 2002, Fehr and Rockenbach 2003).

An interesting issue is to what extent the effect of communication observed in the experiment is dependent on a specific sender of the communication (i.e. whether the content of the communication has an effect of its own, or if the communication has an effect only in a specific relationship such as, here, the dictator-recipient relationship). We call these the content and relationship hypotheses, respectively. They provide two different pathways for how communication may affect behavior in the dictator game. The content of the communication may affect the fairness norm or the cost of deviating from this norm in line with a norm preferences model of the type proposed by Rabin (1994) and Konow (2000). There may also be a relationship specific effect of communication; communication may for instance increase the empathy for the recipient or decrease the social distance between the dictator and the recipient (Bohnet and Frey 1999a, Hoffman et al. 1994, 1996). There is some prior evidence that identification may increase generosity in line with the relationship hypothesis (Bohnet and Frey 1999a, 1999b, Burnham 2003).

To test the content and relationship hypotheses we include a third treatment, with one-way communication from third-parties (in the form of the messages sent from the recipients in the communication treatment). The content hypothesis implies that donations should be higher with third-party communication than with no communication, and the relationship hypothesis implies that donations should be lower with third-party communication than with regular communication. The average donation with third-party communication is about 40 percent higher than with no communication, and this effect is marginally significant ($p<0.10$). The difference between third-party communication and regular communication is not significant, and we cannot reject the null hypothesis. However, the point estimate suggests that the effects of content and relationship are of about the same magnitude.

Section 2 below describes the design of the experiment, and in section 3 we describe the hypotheses to be tested. The results are reported in section 4, and we end with some concluding remarks in section 5.

2. Experimental Design

We carried out three experimental treatments with two sessions of each treatment. The first three sessions (one for each treatment) were carried out in May 2004 and the remaining three sessions were carried out in September 2004.[2] For the May sessions, participants were recruited from undergraduate students at the Stockholm School of Economics, Stockholm University, and the Royal Institute of Technology through e-mail, posters and information in connection with lectures. The participants in the

[2] All sessions were carried out at the Stockholm School of Economics.

September sessions were recently enrolled undergraduate business and economics students at the Stockholm School of Economics. In addition to their earnings in the dictator game, the subjects were paid SEK (Swedish kronor) 50 for participating (the exchange rate at the time of the experiment was: \$1 $\approx$ SEK 7.6). Subjects were randomly allocated to the three treatments.[3]

A total of 348 subjects participated in the experiment. Six of these were used as monitors (see below). The remaining 342 subjects yielded 171 pairs of observations, 57 in each treatment (of these the May and September sessions gave 31 and 26 observations, respectively). In the analysis one pair (with zero donations) was dropped from each session, as one dictator per session received only pieces of paper and no money (as part of the double-blind design, see below). Therefore 55 pairs of observations for each treatment are included in the analysis.

The experiment studies a dictator game, in which the dictator decides how to divide SEK 120 between him/herself and the recipient. In the first treatment there is no communication. In the second treatment the recipients can send a written free-form message to the dictator before the allocation decision takes place. In the third treatment the dictator receives a written message from a third party (a previous recipient) before the allocation decision. The experimental design in all three treatments is double-blind so that neither other subjects nor the experimenters can observe the decision of a particular subject (Hoffman et al. 1994, 1996, Eckel and Grossman 1996, 1998).

In all the three treatments subjects are recruited to two separate rooms called room A and room B.[4] Dictators are in room A and recipients are in room B. The subjects are welcomed and told not to talk to each other. The subjects read the instructions, then the instructions are read aloud by the experimenter, and thereafter the subjects can ask questions individually. The three treatments are further described below.[5]

---

[3] Some subjects could only participate at a specific time and could therefore not be allocated randomly. Such subjects were placed in room B in treatment I or III since these groups had no opportunity to influence the outcome of the experiment. A stratified random selection procedure was used for the May sessions. Subjects were divided into three groups according to their university or school: the Stockholm School of Economics, the Department of Social Work at Stockholm University (since the department is not located on the campus of Stockholm University), and the rest of Stockholm University together with the Royal Institute of Technology. Subjects were also stratified according to gender, which yielded a very equal gender distribution in the three treatments (the number of women among the Dictators was 25 (out of 57) in treatment I, 26 (out of 57) in treatment II, and 26 (out of 57) in treatment III. Significant differences between the sexes have been found in the dictator game by Eckel and Grossman (1998) and Andreoni and Vesterlund (2001).

[4] As pointed out by an anonymous referee it may in general be better to have people all meet initially in one room before separation, so that they see that these other people actually exist. However, the use of a monitor among the students may have mitigated this problem (see below).

[5] The complete instructions can be found in the appendix.

2.1. Treatment I: No communication. A monitor is chosen among the subjects in room A, and he/she conducts the experiment and verifies that the procedures are followed as described in the instructions.[6] The monitor calls one subject at a time in room A and randomly gives the subject an envelope. All envelopes except one contain six SEK 20 bills and six slips of paper of the same size as the bills. The remaining envelope contains twelve slips of paper.[7]

The subject who has received an envelope goes behind a screen. In private behind the screen, the subject removes (and keeps for his/her own use) six units from the envelope (bills or slips of paper), seals the envelope and then drops it in a box marked "Mail".

When all subjects in room A have made their decisions, the monitor brings the box marked "Mail" to room B. The monitor calls one person at a time in room B, opens an envelope, records the contents of the envelope, and gives the contents to the person called. That person can then leave the room. The monitor continues until all envelopes have been opened. The experiment is then over.

2.2. Treatment II: Communication. The communication treatment is carried out in a similar way as the "no communication treatment". The difference is that before the dictator makes the allocation decision, the subjects in room B are given ten minutes to write a message on a numbered "message form" received together with the instructions.[8] Each subject in room A then reads one of these messages prior to making their allocation decision.

2.3. Treatment III: Third-party communication. The third-party communication treatment is also carried out in a similar way as the "no communication treatment". The difference is that each subject in room A receives and reads a message prior to making his/her allocation decision. The message is a message from a previous recipient in treatment II, and in the instructions the subjects are informed that this is a message sent from a recipient to a dictator in a previous experiment the same day.

3. Hypotheses and tests

3.1. Hypotheses. Let μ_1, μ_2, and μ_3 denote the mean donation in treatments I, II, and III, respectively. We test three hypotheses.

[6] The monitor receives SEK 120 in addition to the SEK 50 already received.

[7] The envelope without money is included as an additional guarantee of anonymity. Even if no dictator donates any money the experimenter will not be able to infer the decision of a single subject.

[8] As the experiment now "starts" in room B, the monitor is chosen among the subjects in room B rather than in room A.

3.1.1. *Communication hypothesis.* Our first hypothesis to be tested is that there is an effect of anonymous communication. It implies

Hypothesis 1: Donations are higher with communication (Treatment II) than without communication (Treatment I); that is $\mu_2 > \mu_1$.

In this paper the effect of communication is divided into the content effect and the relationship effect. This hypothesis can therefore also be interpreted as a test of the joint effect of the content and relationship effects of communication (tested separately in hypotheses 2 and 3).

3.1.2. *Content hypothesis.* We define the content hypothesis of communication in the following way. The content of communication has an effect on behavior that is not dependent on a specific sender of the communication. This hypothesis implies

Hypothesis 2: Donations are higher with third-party communication (Treatment III) than without communication (Treatment I); that is $\mu_3 > \mu_1$.

3.1.3. *Relationship hypothesis.* We define the relationship hypothesis in the following way. The content of communication has an effect on behavior that is dependent on a specific sender. This hypothesis implies

Hypothesis 3: Donations are higher with communication (Treatment II) than with third-party communication (Treatment III); that is $\mu_2 > \mu_3$.

3.2. Statistics. To compare average donations between treatments we use bootstrap techniques, because bargaining experiments usually lead to skewed distributions.[9] Bootstrap techniques make it possible to conduct statistical testing without imposing normality (i.e. by inferring the underlying distribution from which the data has emerged).[10] Ellingsen and Johannesson discuss the choice of test statistics in more detail.

The significance levels of comparisons of average donations we report below have all been obtained by generating 2,099 bootstrap replications. According to Davidson and MacKinnon (2000), this number of replications is high enough to guarantee a reasonable confidence in the estimated p-values, compared to the "ideal" bootstrap with infinitely many replications. For comparison we also report the significance level with the nonparametric Mann-Whitney test, which is commonly used to analyse experimental data.

[9] This is also the case in our data. According to a Kolmogorov-Smirnov test, normality can be rejected in all the three treatments at the 1% level.

[10] For an introduction to bootstrap methods, see for example Efron and Tibshirani (1993).

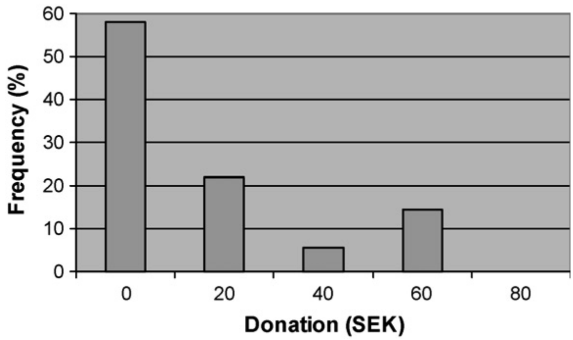

FIGURE 1. Distribution of donations in Treatment I (no communication).

Since the hypotheses are directional we use one-tailed tests and all reported p-values are one-sided.

4. Results

The distributions of donations in the three treatments are shown in Figures 1-3. Table 1 provides descriptive results, and the test results are provided in Table 2.

In the treatment with no communication 42% of subjects (dictators) donate some money, and the most common donation is SEK 20 (the lowest possible donation). The average donation is 12.73% of the endowment. This is very similar to a previous double-blind dictator game experiment on student subjects at the Stockholm School of Economics, which yielded an average donation of 13.33% (Johannesson and Persson 2000). It is also similar to previous studies in the US with a similar double-blind experimental design, where the average donation has ranged between 8% and 16% of the amount allocated (Hoffman et al. 1994, 1996, Eckel and Grossman, 1996, 1998, Burnham 2003).

4.1. Hypothesis 1: Communication hypothesis. With communication the share of subjects who donate some money increases from 42% to 58%. The most common donation is the equal split, which is made by 53% of donors (compared to 35% of donors with no communication). The average donation increases from 12.73% to 22.13%, and this difference is significant at the 5% level (according to both the bootstrap and the Mann-Whitney test). Our first hypothesis that there is an effect of communication is therefore supported. The effect of communication is sizeable, leading to an increase in average donations of more than 70%.

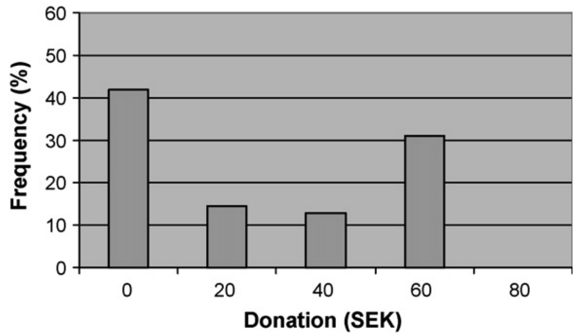

FIGURE 2. Distribution of donations in Treatment II (communication).

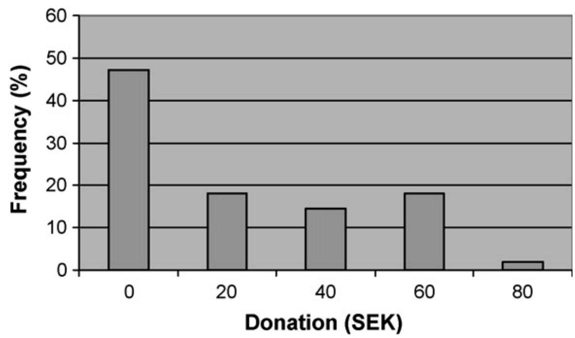

FIGURE 3. Distribution of donations in Treatment III (third-party communication).

4.2. Hypothesis 2: Content hypothesis. With third party communication the share of subjects who donate some money increases from 42% to 53% compared to no communication. The equal split is made by 34% of donors, which is similar to the no communication treatment. The average donation increases from 12.73% to 18.18%, an increase of over 40%. This difference is marginally significant (p=0.065 with the bootstrap test and p=0.083 with the Mann-Whitney test), supporting the content hypothesis.

4.3. Hypothesis 3: Relationship hypothesis. The share of subjects who donate some money is somewhat higher with communication than with third-party communication, 58% versus 53%. The equal split is also more common among donors with communication than with third-party communication, 53% versus 34%. The average

TABLE 1. Experimental results

Amount donated:	Treatment: (I) No Communication	(II) Communication	(III) Third-Party Communication
SEK 0	32	23	26
SEK 20	12	8	10
SEK 40	3	7	8
SEK 60	8	17	10
SEK 80	0	0	1
Number of observations	55	55	55
Average donation in SEK	15.27 (21.76)	26.55 (26.12)	21.82 (24.73)
Average donation in percent	12.73 (18.13)	22.13 (21.77)	18.18 (20.61)

Standard deviation in parenthesis

donation is 22.13% with communication compared to 18.18% with third-party communication, but this difference is not significant (p=0.137 according to the bootstrap test and p=0.175 according to the Mann-Whitney test). Therefore, we cannot reject the null hypothesis of no relationship effect of communication.

TABLE 2. Tests of differences in average donations between treatments

	Bootstrap	Mann-Whitney
No communication versus communication (Hypothesis 1)	0.006	0.013
No communication versus third-party communication (Hypothesis 2)	0.065	0.083
Communication versus third-party communication (Hypothesis 3)	0.137	0.175

Numbers are one-sided p-values for tests of treatment differences in average donations.

5. Concluding remarks

Communication significantly increased donations in our experiment, and we could reject the null hypothesis of no effect of communication. The effect of communication was also relatively large, increasing donations by more than 70%.

We also attempted to estimate how much of the effect of communication is due to the content of the communication per se and how much is "relationship specific". This was achieved by adding an experimental treatment with third-party communication (in the form of messages from the previous treatment). Average donations in this treatment were more than 40% higher than with no communication, and the effect was marginally significant ($p<0.10$). This provides support for the hypothesis that the content of the communication per se affects donations. For the relationship hypothesis we failed to reject the null hypothesis, but the point estimate suggests that the content and relationship effects are of about the same magnitude.

To gain some further insights into the communication process we read all the 57 messages and attempted to classify them into different categories. The content of the messages is consistent with both the content and relationship hypotheses. A majority of subjects (72%) provided impersonal arguments for giving, and these arguments typically appealed to some kind of moral reasonsfor giving such as goodness, fairness, or avoiding a bad conscience. The frequent use of these impersonal arguments provides a scope for third party communication to affect donations in line with the content hypothesis. About a third of the subjects (35%) provided personal arguments for giving, and the most common personal argument was to refer to the recipient's great need or low wealth. About half the subjects (53%) also provided non-argumentative contents such as cordial greetings or some information about the identity of the recipient. The personal arguments and the communication without argumentative content should be of less importance in third-party communication, but may increase the sympathy between the communicating persons or reduce social distance in line with the relationship hypothesis. We tried to analyse the effectiveness of different type of messages in a regression analysis, but found no consistent differences between different types of messages; however the sample may be too small to detect differences between different types of messages.

Our work is related to the work on identification in the dictator game by Bohnet and Frey (1999a, 1999b). They found that one-way visual identification, where dictators could see recipients, increased donations by 35%. When the one-way visual identification was combined with information about the recipient (their name, hobbies, major and where they came from) donations approximately doubled compared to no identification. Burnham also carried out a dictator game with one-way visual identification, where the dictators were shown a photo of the recipient prior to the allocation decision. This increased donations by about 65% compared to no identification. Although one-way identification is not the same as communication, the results of Bohnet and Frey (1999a, 1999b) and Burnham suggest that more information about the identity of the recipient can increase donations (in line with the relationship hypothesis). However, Dufwenberg and Muren (2006) found a result that in a sense runs in the opposite direction. When they decreased the anonymity of the dictator, by having the dictator come on stage to receive the payment, donations decreased.

Frey and Bohnet (1995) and Bohnet and Frey (1999b) furthermore found that mutual silent visual identification in the dictator game nearly doubled donations. Those results are less comparable to our results since mutual identification creates possibilities for dictators to engage in reputation building behavior. This is not the case when the communication or identification is one-way and dictators remain anonymous.

If players are aware of the effect of communication documented here, the question arises as to what extent communication is used strategically to increase other people's other-regarding behavior towards them and to decrease their own other-regarding behavior towards others. Empirically, Frey and Bohnet tested whether subjects in a dictator game wanted to communicate face-to-face or not. When communication was free, 86% of the recipients and 75% of the dictators chose to communicate. The difference in the willingness to communicate between dictators and recipients may be due to dictators realizing that they will contribute more if they communicate. But it is still difficult to understand why so many dictators choose to communicate, although they may be driven by curiosity or the desire to engage in reputation building.

Theoretically, little work has been done on endogenous communication. However, the model of guilt aversion of Charness and Dufwenberg (2006) provides a rationale for endogenous communication. According to guilt aversion a player suffers from guilt if she hurts others relative to what they believe they will get. The dictator is thus motivated by the beliefs about the beliefs of the recipient. Communication may affect these beliefs and thereby affect allocations. This model provides an interesting interpretation of the effect of communication in our experiment. The relationship specific effect of communication in the experiment is consistent with guilt aversion as the message may provide information about the beliefs of the recipient. It is less straightforward to explain the observed sender-independent effect of communication with guilt aversion, but it is possible that also sender-independent communication may be used to update beliefs about the recipient as it may provide information about the average beliefs in the population.

In ending, we conclude that anonymous communication increases donations in dictator games. The results also suggest that the content of communication causes an effect that is independent of the sender. Further work is needed to pin down the content and relationship effects of communication with greater precision.

Appendix: Experimental Instructions

The original instructions were in Swedish. This appendix reprints a translation of the instructions used in the three treatments (no communication, communication, and third-party communication).

A1. Instructions (Treatment I: No communication). Thank you for participating in this experiment. For your participation you have received SEK 50. In addition to this you have the opportunity to receive more money during the experiment (maximally SEK 120).

In this experiment each of you will be paired with another person in another room. You will not get to know who these persons in the other room are, neither during, nor after the experiment. Except for one person in room A, who will be chosen to be a monitor, there is an equal number of persons in each room (A and B). **This is room A (B)**. Every person in room A and room B has received these instructions as well as SEK 50 for participating in the experiment. In the experiment each person in room A (except for the monitor and one other person, see below) will decide how to divide SEK 120 between him/herself and the person in room B with whom he/she has been paired.

One of the persons in room A will be chosen to monitor the experiment. The monitor will get SEK 120 in addition to the SEK 50 that person has already received. The monitor's task is to take care of the envelopes we will describe soon. Furthermore the monitor shall control and certify that the instructions we now go through were followed.

The experiment runs as follows. Unmarked envelopes corresponding to the number of participants have been put into a box in room A. All except one of these contain six SEK 20 bills and six blank pieces of paper of the same size. The remaining envelope contains 12 blank pieces of paper. The monitor calls one person at a time in room A and gives that person one of the envelopes from the box. The person takes the envelope with him/her and goes behind the screen in room A. The envelope is then opened behind the screen where no one can see what happens.

Behind the screen every person in room A has to decide how many bills and how many pieces of paper to leave in the envelope. The number of bills and pieces of paper left in the envelope shall be six in total. The person then pockets the remaining pieces of paper and bills. Example: (1) The person leaves SEK 20 and five pieces of paper in the envelope and keeps SEK 100 and one piece of paper for him/herself; (2) The person leaves SEK 80 and two pieces of paper in the envelope and keeps SEK 40 and four pieces of paper for him/herself. These were only examples and the real decision is up to each person in room A. No one else, including those conducting the experiment, will know what decision a particular person makes.

When the person behind the screen has made his/her decision he/she seals the envelope and puts it in the box marked “Mail”. The person can then leave the room.

When all envelopes have been handed in the monitor takes the box with envelopes to room B. The monitor asks one person at a time in room B to come forward. Then the monitor takes up an envelope from the box, opens it and writes down the contents and then gives the contents to the person in question. That person can then leave the room. The monitor continues until all envelopes have been opened and everyone has left the room. The experiment is then over.

A2. Instructions (Treatment II: Communication). Thank you for participating in this experiment. For your participation you have received SEK 50. In addition to this you have the opportunity to receive more money during the experiment (maximally SEK 120).

In this experiment each of you will be paired with another person in another room. You will not get to know who these persons in the other room are, neither during, nor after the experiment. Except for one person in room A, who will be chosen to be a monitor, there is an equal number of persons in each room (A and B). **This is room A (B)**. Every person in room A and room B has received these instructions as well as SEK 50 for participating in the experiment. In the experiment each person in room A (except for the monitor and one other person, see below) will decide how to divide SEK 120 between him/herself and the person in room B with whom he/she has been paired.

One of the persons in room A will be chosen to monitor the experiment. The monitor will get SEK 120 in addition to the SEK 50 that person has already received. The monitor’s task is to take care of the envelopes we will describe soon. Furthermore the monitor shall control and certify that the instructions we now go through were followed.

The experiment runs as follows. In room B each person has received a number and a sheet of paper marked “message” (which also has the same number). The persons in room B are given 10 minutes to write a message to the person in room A. After 10 minutes the monitor collects these messages and puts each message into one unmarked large envelope. In every large envelope there also lies a smaller envelope. All except one of these smaller envelopes contain six SEK 20 bills and six blank pieces of paper of the same size. The remaining smaller envelope contains 12 blank pieces of paper. The monitor then takes the envelopes (with messages and money/pieces of paper) to room A, and gives one envelope to each person in room A.

When the persons in room A have gotten their envelopes they open the large envelope and take out the paper marked "message", but let the smaller envelope remain inside the large envelope. Each person reads his/her message silently. Thereafter everyone puts the message back into the envelope.

The monitor then asks one person at a time in room A to come forward. The person takes the envelope and goes behind the screen in room A. The smaller envelope with the money/pieces of paper is then opened behind the screen in room A where no one can see what happens.

Behind the screen every person in room A has to decide how many bills and how many pieces of paper to leave in the smaller envelope. The number of bills and pieces of paper left in the envelope shall be six in total. The person then pockets the remaining pieces of paper and bills. Example: (1) The person leaves SEK 20 and five pieces of paper in the envelope and keeps SEK 100 and one piece of paper for him/herself; (2) The person leaves SEK 80 and two pieces of paper in the envelope and keeps SEK 40 and four pieces of paper for him/herself. These were only examples and the real decision is up to each person in room A. No one else, including those conducting the experiment, will know what decision a particular person makes.

When the person behind the screen has made his/her decision he/she seals the smaller envelope and puts it in the larger envelope which is also sealed. The person then puts the envelope in the box marked "Mail". The person can then leave the room.

When all envelopes have been handed in the monitor takes the box with envelopes to room B. The monitor takes up an envelope from the box, opens it and writes down the contents of the smaller envelope. The monitor then asks the person with the number on the form marked "message" to come forward and gives the contents of the smaller envelope to him/her. That person can then leave the room. The monitor continues until all envelopes have been opened and everyone has left the room. The experiment is then over.

A3. Instructions (Treatment III: Third-party communication). Thank you for participating in this experiment. For your participation you have received SEK 50. In addition to this you have the opportunity to receive more money during the experiment (maximally SEK 120).

In this experiment each of you will be paired with another person in another room. You will not get to know who these persons in the other room are, neither during, nor after the experiment. Except for one person in room A, who will be chosen to be a monitor, there is an equal number of persons in each room (A and B). **This is room A (B)**. Every person in room A and room B has received these instructions as well as SEK 50 for participating in the experiment. In the experiment each person in room

A (except for the monitor and one other person, see below) will decide how to divide SEK 120 between him/herself and the person in room B with whom he/she has been paired.

One of the persons in room A will be chosen to monitor the experiment. The monitor will get SEK 120 in addition to the SEK 50 that person has already received. The monitor's task is to take care of the envelopes we will describe soon. Furthermore the monitor shall control and certify that the instructions we now go through were followed.

The experiment runs in the following way. Unmarked large envelopes corresponding to the number of participants have been put into a box in room A. In each of these envelopes there is a smaller envelope as well as a sheet of paper marked "message" (these messages are described below). All except one of the smaller envelopes contain six SEK 20 bills and six blank pieces of paper of the same size. The remaining smaller envelope contains 12 blank pieces of paper.

The monitor hands out an envelope to each person in room A. When the persons in room A have gotten their envelopes they open the large envelope and take out the form marked "message", but let the smaller envelope remain inside the large envelope. Each person reads his/her message silently. Thereafter everyone puts the message back into the envelope.

The monitor then asks one person at a time in room A to come forward. The person takes the envelope and goes behind the screen in room A. The smaller envelope with the money/pieces of paper is then opened behind the screen in room A where no one can see what happens.

Behind the screen every person in room A has to decide how many bills and how many pieces of paper to leave in the smaller envelope. The number of bills and pieces of paper left in the envelope shall be six in total. The person then pockets the remaining pieces of paper and bills. Example: (1) The person leaves SEK 20 and five pieces of paper in the envelope and keeps SEK 100 and one piece of paper for him/herself; (2) The person leaves SEK 80 and two pieces of paper in the envelope and keeps SEK 40 and four pieces of paper for him/herself. These were only examples and the real decision is up to each person in room A. No one else, including those conducting the experiment, will know what decision a particular person makes.

When the person behind the screen has made his/her decision he/she seals the smaller envelope and puts it in the larger envelope which is also sealed. The person then puts the envelope in the box marked "Mail". The person can then leave the room.

When all envelopes have been handed in the monitor takes the box with envelopes to room B. The monitor asks one person at a time in room B to come forward. Then the monitor takes up an envelope from the box, opens it and writes down the contents and then gives the contents to the person in question. That person can then leave the room. The monitor continues until all envelopes have been opened and everyone has left the room. The experiment is then over.

The paper marked ”message” comes from an earlier experiment today. The earlier experiment was identical to this experiment, except that every person in room B then could send a message to the person in room A before that person decided how to divide the SEK 120. In every envelope which is handed out to the persons in room A in this experiment we have put a message from the earlier experiment. The persons who now sit in room B have not had the possibility to send any messages.

References

Andreoni, J., Vesterlund, L., (2001), 'Which is the fair sex? Gender differences in altruis', *Quarterly Journal of Economics* 116, 293-312.

Bochet, O., Page, T., Putterman, L., (2002), 'Communication and punishment in voluntary contribution experiments', Working Paper 2002-29, Brown University.

Bohnet, I., Frey, B.S., (1999a), 'The sound of silence in prisoner's dilemma and dictator games' *Journal of Economic Behavior and Organization* 38, 43-57.

Bohnet, I., Frey, B.S., (1999b), 'Social distance and other regarding behavior in dictator games: comment', *American Economic Review* 89, 335-339.

Bolton, G.E., Ockenfels, A., (2000), 'ERC: a theory of equity, reciprocity, and competition', *American Economic Review* 90, 166-193.

Brosig J., Weimann, J., Yang, C-L., (2004), 'Communication, reputation, and punishment in simple sequential bargaining experiments', *Journal of Institutional and Theoretical Economics* 160, 576-606.

Burnham, T.C., (2003), 'Engineering altruism: a theoretical and experimental investigation of anonymity and gift giving', *Journal of Economic Behavior and Organization* 50, 133-144.

Camerer, C.F., (2003), *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton: Princeton University Press.

Charness, G., Dufwenberg, M., (2006), 'Promises and partnership', *Econometrica* 74(6), 1579-1601.

Charness, G., Rabin, M., (2002), 'Understanding social preferences with simple tests', *Quarterly Journal of Economics* 117, 817-869.

Charness, G., Rabin, M., (2005), 'Expressed preferences and behavior in experimental games', *Games and Economic Behavior* 53, 151-169.

Crawford, V., (1998), 'A survey of experiments on communication via cheap talk', *Journal of Economic Theory* 78, 286-298.

Davidson, R., MacKinnon, R.G., (2000), 'Bootstrap tests: how many bootstraps?', *Econometric Reviews* 19, 55-68.

Dufwenberg, M., Muren, A., (2006), 'Generosity, anonymity, gender', *Journal of Economic Behavior and Organization* 61, 42–49.

Eckel, C.C., Grossman, P.J., (1996), 'Altruism in anonymous dictator games', *Games and Economic Behavior* 16, 181-191.

Eckel, C.C., Grossman, P.J., (1998), 'Are women less selfish than men?: Evidence from dictator experiments', *Economic Journal* 108, 726-735.

Efron, B., Tibshirani, R.J., (1993), 'An Introduction to the Bootstrap', *Monographs on Statistics and Applied Probability* 57. New York: Chapman and Hall.

Ellingsen, T., Johannesson, M., (2004), 'Promises, threats, and fairness', *Economic Journal* 114, 397-420.

Fehr, E., Gächter, S., (2002), 'Do incentive contracts undermine voluntary cooperation?' Working Paper No. 34, Institute for Empirical Research in Economics, University of Zurich.

Fehr, E., Klein, A., Schmidt, K., (2001), 'Fairness, incentives and contractual incompleteness' Working Paper No. 72, Institute for Empirical Research in Economics, University of Zurich.

Fehr, E., Rockenbach, B., (2003), 'Detrimental effects of sanctions on human altruism', *Nature* 422, 137-140.

Fehr, E., Schmidt, K., (1999), 'A theory of fairness, competition and cooperation', *Quarterly Journal of Economics* 114, 817-868.

Frey, B.S., Bohnet, I., (1995), 'Institutions affect fairness: experimental investigations', Journal of Institutional and Theoretical Economics 151, 286-303.

Frohlich, N., Oppenheimer, J., (1998), 'Some consequences of e-mail vs. face-to-face communication in experiment', *Journal of Economic Behavior and Organization* 35, 389-403.

Hoffman, E., McCabe, K., Shachat, K., Smith, V.L., (1994), 'Preferences, property rights, and anonymity in bargaining games', *Games and Economic Behavior* 7, 346 -380.

Hoffman, E., McCabe, K., Smith, V.L., (1996), 'Social distance and other regarding behavior in dictator games', *American Economic Review* 86, 653 -660.

Johannesson, M., Persson, B., (2000), 'Non-reciprocal altruism in dictator games', Economics Letters 69, 137-142.

Konow, J., (2000), 'Fair shares: accountability and cognitive dissonance in allocation decisions', *American Economic Review* 90, 1072-1091.

Ledyard, J.O., (1995), 'Public goods: a survey of experimental research', In: Kagel, J.H., Roth, A.E. (Eds.). *Handbook of Experimental Economics.* Princeton: Princeton University Press, 111-194.

Rabin, M., (1993), 'Incorporating fairness into game theory and economics', *American Economic Review* 83, 1281-1302.

Rabin, M., (1994), 'Cognitive dissonance and social change', *Journal of Economic Behavior and Organization* 23, 177-194.

Sally, D., (1995), 'Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958-1992', *Rationality and Society* 7, 58-92.

Sen, A.K., (1967), 'Isolation, assurance and the social rate of discount', *Quarterly Journal of Economics* 81, 112-124.

Valley, K., Thompson, L., Gibbons, R., Bazerman, M., (2002), 'How communication improves efficiency in bargaining games', *Games and Economic Behavior* 38, 127-155.

PAPER 2

Limbic Justice – Amygdala Drives Rejection in the Ultimatum Game

Katarina Gospic*, Erik Mohlin*, Peter Fransson, Predrag Petrovic, Magnus Johannesson, and Martin Ingvar

ABSTRACT. The ultimatum game is a stylized game to study decision making in which a proposer suggests how to split a sum of money either fairly or unfairly. Unfair splits are often rejected by the responder even at a personal cost. Previous research using unspecific stimulus-onsets suggests that such rejections are of cortical origin and involve a change of feeling states. Using an fMRI-design to specifically study early emotional components of decision making and a pharmacological intervention we demonstrate a causal role of the limbic system in the act of rejection. In the placebo-treated group rejection was directly linked to an increased amygdala activity and benzodiazepine treatment decreased rejection rate concomitantly with a suppressed amygdala response to unfair proposals in spite of an unchanged feeling of unfairness. Thus, we segregate the neural basis of rejections associated with the initial reactive emotional response from the slower affective processing associated with awareness.

1. Introduction

Research within behavioral economics and psychology has demonstrated that human decisions are based on more dimensions than simply maximization of monetary reward (Tversky and Kahneman 1981, De Martino et al. 2006, Camerer 2003). One important factor with prominent impact on decision making is the influence of emotions (Bechara et al. 2003). Emotional responses are rapid and automatic in order to meet the demands for fast contextual adaptation. On the other hand, the representation of feeling states and the regulatory control of emotions reflects a slower adjustment to long term considerations and goals (Craig 2009). A human universal in social cooperation is

---

* Katarina Gospic and Erik Mohlin contributed equally to this paper.
K.G., E.M., P.F., M.J. and M.I designed the experiment. K.G. and E.M. conducted the experiment. K.G., P.F., P.P. and M.I. analyzed the fMRI data and E.M. analyzed the behavioural data. K.G., E.M., P.F., P.P., M.J. and M.I wrote the paper.
This work was funded by the Swedish Research Council, The Barbro and Bernard Osher Foundation, The Swedish Agency for Innovation Systems (VINNOVA), The Swedish Foundation for Strategic Research, The Jan Wallander and Tom Hedelius Foundation, The Swedish Council for Working Life and Social Research, The Knut and Alice Wallenberg Foundation and the Karolinska Institute.

the tendency to respond with immediate aggression upon perceived threat or unfairness (Pinker 2003). Evolution seems to have premiered the act of punishing those who violate perceived norms of the group (Fehr and Gächter 2000). Recently, brain imaging studies have shown that emotional systems are active in decision making (Sanfey et al.2003, Tricomi et al.2010, Buckholtz et al.2008). However, as these studies are purely correlational the causal role of the different sub-regions involved in the decision making network is yet to be established.

A suitable paradigm to study the punishment of norm violating behavior is the Ultimatum Game (UG) (Güth et al. 1982). In the UG, a proposer suggests a way to divide a fixed sum of money. The responder has to accept or reject the proposal. If the responder accepts the proposal, the suggested split is realized. If the responder rejects the offer, both subjects get nothing. The proposals can either be of fair (50/50) or unfair nature (e.g. 20/80). Unfair offers are frequently rejected, and offers below 20% are rejected, roughly half of the time. These findings are robust with respect to learning effects, stake size, and other manipulations (Camerer 2003). Although both individual genetic traits and cultural variation influence the response pattern the general propensity to punish norm violators seems to be universal (Wallace et al. 2007, Henrich et al. 2005).

In the UG the payoff-maximizing strategy for the responder is to accept all offers and reciprocally for the proposer to make the smallest possible offer (Güth et al. 1982). Several studies suggest that emotional theory may add important information for choice behavior (Tversky and Kahneman 1981, De Martino et al. 2006). The UG demands a simple, rapid yes or no answer but the underlying reasons for response are complex and have both short and long term perspectives. A short term reason for instant rejection of an unfair proposal could be that the perceived unfairness invokes an automatic reactive aggressive response (i.e. a tit-for-tat response). The short term emotional reaction could alternatively bias decisions towards acceptance by means of reward expectation mechanisms or because the perceived social rank of the proposer is superior and thereby perceived as an implicit threat. A more long term reason for rejection could be to maintain social norms, whereas a long term bias towards acceptance lies in maximizing the monetary reward. The short term unaware responses are instantiated in the subcortical emotion system (e.g. amygdala) (Bechara et al. 2003, Vuilleumier 2005) whereas the long term considerations pertain frontal cortex and insula (Bechara et al.1998, Xue et al. 2010). Thus, the response in the UG rests on a balance between phylogenetically ancient structures involved in the automatic emotional response vs. neocortical areas associated with the neural processing of feeling states and regulation of emotions (Craig 2009, Damasio 1994).

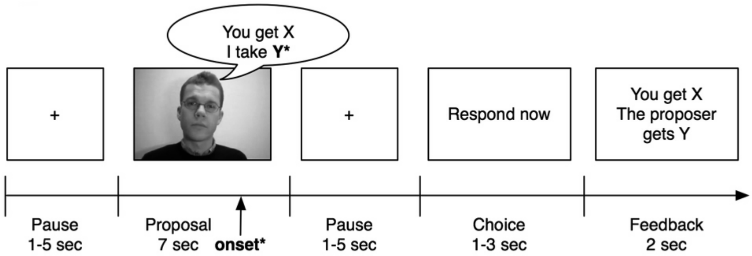

FIGURE A1. *Experimental set-up.* 35 subjects were randomly assigned to either the control or the treatment group (oxazepam 20 mg p.o.). One hour after treatment subjects played the ultimatum game in the scanner by watching 45 movie clips, each with a different human proposer. The proposals were either fair, unfair, or neutral. All proposals had the exact same wording and the proposer ended the sentence with stating the share that he/she would get. The fMRI onset time was defined as to when the last word was spoken i.e. when the fairness of the proposal could be judged. Subjects were instructed to respond either yes or no to the fair/unfair proposals and no to the neutral proposals. Post scanning, subjects rated the fairness of the offers (scale 1-7) (8) and likeability of the proposers faces (scale 0-100).

Previous imaging studies (Sanfey et al.2003, Rilling et al.2004) on decision making in the UG has shown increased neural activity in a forebrain network (anterior insula, dorsolateral prefrontal cortex (dlPFC), and anterior cingulate cortex (ACC)) in response to unfair UG offers. The authors suggested that these structures were involved in the neural processing of perceived unfairness, thereby driving the behavior to reject unfair offers. However, those studies did not attempt to separate the instant automatic responses from the slower affective processes associated with awareness (Craig 2009). The immediate responses are likely to be transient and mitigated when emotional regulation sets in (Vuilleumier et al. 2001). Thus, a prerequisite to detect these responses is that the onset time is strictly defined for when unfairness is elicited. Fast automatic responses in the UG have not previously been captured since the proposals were presented for 6 seconds (Sanfey et al.2003, Rilling et al.2004) thereby removing a clear definition of the onset time. In our experiment, the proposals were given orally in movie clips and we formulated the UG proposal as to maintain ambiguity of fairness until the final word of the amount that would be taken by the proposer was spoken. Hence, the onset time was well defined in the present experiment (Fig. 1).

As amygdala is crucial for both the mediation of aggressive responses (Nelson and Trainor 2007) and of biasing decision making (De Martino et al. 2006, Bechara et al.

2003) we suggest a parallel between reactive aggression and the behavior associated to rejection. Thus, we hypothesized that amygdala drives immediate rejection in the UG.

GABA receptors are abundant in the amygdala and benzodiazepines can potentiate GABA activity, reduce behavioral signs of aggression (Nelson and Trainor 2007), and decrease amygdala activity in emotional tasks (Arce et al. 2006, Paulus et al. 2005). In the present study, we posited that the benzodiazepine oxazepam (20 mg p.o.) could inhibit amygdala activity and thus, change behavior in the UG. We assumed that unfair proposals would generate an amygdala response and increase rejection rate in the non-medicated group while oxazepam would inhibit this process and therefore, reduce the rejection rate in response to unfair proposals in parallel with a mitigated amygdala response to unfair proposals.

2. Results

2.1. Treatment decreases rejection rate of unfair proposals. In line with previous studies, fair proposals were never rejected in either group (Sanfey et al.2003, 23). The rejection rate for unfair proposals was significantly lower in the oxazepam group (19%) (n=17) compared to the placebo group (37.6%) (n=18) (Fig. 2A). The fMRI contrast unfair vs. fair in the placebo group essentially confirmed the results from Sanfey et al. 2003, Fig. 3A & 3B. The corresponding contrast in the oxazepam group showed a subsignificant activation in the right insula (see supporting information Fig. S2).

2.2. Treatment suppresses neural activity related to rejection. Given that oxazepam inhibits rejection of unfair offers the interesting contrast is the interaction: placebo $_{\text{unfair – fair proposals}}$ > oxazepam $_{\text{unfair - fair proposals}}$. We confirmed our primary hypothesis that amygdala was relatively more activated in the placebo group than in the oxazepam group for unfair offers (left amygdala: [-18 -6 -18] $Z = 3.25$, $p = 0.05$ corrected; right amygdala: [18 0 -18] Z score $= 3.03$, $p < 0.05$ corrected) (Fig. 3C). Moreover, the extended fMRI analysis revealed interaction differences in mPFC ([-6 66 18] $Z = 3.77$, $p < 0.05$ cluster level corrected) and right ACC ([9 48 24] $Z = 3.57$, $p < 0.05$ cluster level corrected). Thus, subjects who were treated with oxazepam had a lower activity in specific components of the neural network elicited by unfair proposals (Sanfey et al. 2003) and rejection behavior. Importantly no effects related to rejection were observed in dlPFC and anterior insula seemingly in contrast to Knoch et al. (2006) and Sanfey et al. (2003).

2.3. Rejection of unfair proposals increases amygdala activity in the placebo group. We predicted that increased amygdala activity would correspond

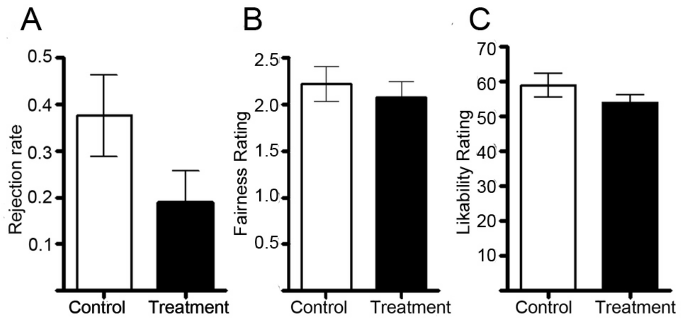

FIGURE A2. *Rejection rate and subjective ratings of fairness/likeability.* (A) Treatment with oxazepam (n = 17) reduced the rejection rate of unfair offers by 49% compared to the control treatment (n = 18) (Mann-Whitney U-test, one-tailed, Z = 1.722, p = 0.049) (B) Rating of fairness for unfair offers and (C) likeability rating of the proposers for rejected/accepted proposals did not change with treatment (Fairness; Z = 0.658, p = 0.51, Likeability; Z = 0.603, p = 0.55). All the ratings were analyzed with the Mann-Whitney U test, two-tailed.

to an increased rejection rate also within the placebo group, as amygdala is known to have a crucial role in decision making (De Martino et al. 2006, Bechara et al. 2003) and aggression (Nelson and Trainor 2007). To test this hypothesis we did a within subject analysis in placebo subjects that both accepted and rejected unfair proposals (n=6). The contrast unfair proposals $_{\text{rejected}}$ > unfair proposals $_{\text{accepted}}$ showed increased amygdala activity ([21 -3 -12]; Z score = 3.50, p < 0.05 corrected) (Fig. 4A).

2.4. Males show a greater amygdala response to unfair proposals. As testosterone can increase aggressive behavior (Nelson and Trainor 2007) we tested whether males drive the reactive amygdala response and hence would show an increased amygdala activity in response to unfairness. To test this hypothesis we compared amygdala activity between sexes for the contrast unfair > fair proposals. Strikingly, males (n=5) showed a greater right amygdala activity compared to females (n=12) in the placebo condition while there was no difference between sexes in the oxazepam condition (males n=8, females n=10) (Fig. 4B). Hence, the interaction sex x treatment was significant (F (1) = 8.50, p = 0.007).

2.5. No effects of treatment on ratings of unfairness and likeability. The subjects treated with oxazepam displayed a decreased rejection rate to unfair proposals

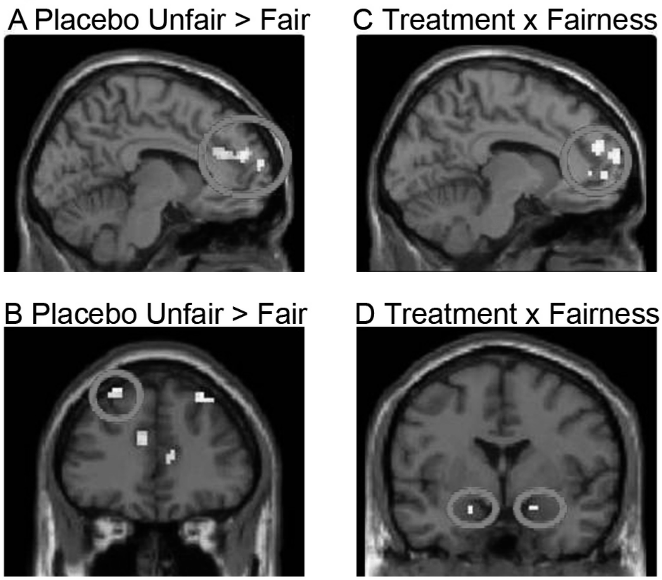

FIGURE A3. *fMRI data related to unfair proposals.* (A) In the placebo group (n = 18) we replicated data from Sanfey et al., 2003 in that unfair proposals elicited activity in right ACC ([9 48 24]; Z=3.15, p<0.001 uncorrected) and (B) bilateral dlPFC (MNI space coordinates (x,y,z): left [-24 36 54]; Z=4.04, p < 0.001 uncorrected, right [30 36 51]; Z=4.04, p>0.001 uncorrected). (C) Here, we show more expressed responses in the placebo group (interaction placebo unfair-fair proposals > oxazepam unfair-fair proposals) with an increased activation in left mPFC ([-6 66 18] Z=3.77, p < 0.05 cluster level corrected), ACC ([9 48 24] Z= 3.57, p < 0.05 cluster level corrected), and (d) bilateral amygdala (left: [-18 -6 -18] Z= 3.25, p < 0.05 corrected; right: [18 0 -18] Z= 3.03, p < 0.05 corrected). Treatment with oxazepam (n = 17) lowered the neural responses related to unfair proposals.

(Fig. 2A). In order to probe the possibility that changed behavior was due to drug-altered perception of unfairness or likeability of the proposer we compared subjective ratings between groups. Subjects in the oxazepam group had similar perception of unfairness (Mann-Whitney U test, two-tailed, p=0.5103) (Fig. 2B) and perceived likeability (Mann-Whitney U test, two-tailed, Z=0.63, p=0.5467) of the proposers as in the control group (Fig. 2C). This is in concordance with the finding that there was no difference in the insula activity between the groups in unfair vs. fair offers (see supporting information), as insula is involved in the coding of feeling states (Craig 2009).

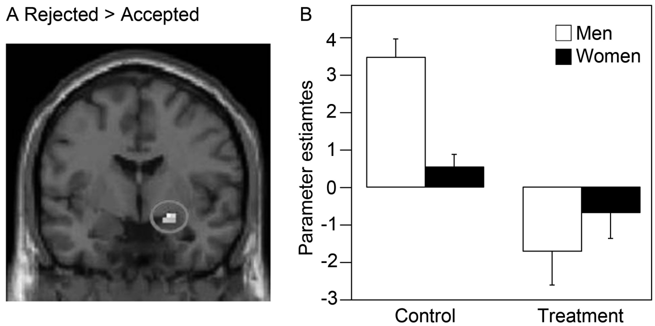

FIGURE A4. *fMRI results related to rejection of an unfair proposal and sex difference.* (A) The rejection of an unfair proposal was associated with a higher activity in the right amygdala ([21 -3 -12]; Z = 3.50, $p < 0.05$ corrected) in the placebo group (n = 6). (B) The difference between sexes for unfair vs. fair proposals in the placebo condition was significant (t (15) = 4.30, two-tailed, p = 0.001) (males n = 5, females n = 12). There was also a main effect of treatment (F (1) = 18.53, p = 0.000) and an interaction between gender x treatment for unfair vs. fair proposals (F (1) = 8.50, p = 0.007). Mean ± S.E.M. of parameter estimates in the right amygdala.

Thus, we found that the observed change in choice behavior between the treatment groups was not explained by an altered feeling of unfairness or insula activity.

3. Discussion

Our study shows that emotional processes are causally involved in choice behavior. Moreover, we segregated the subcortical network that mediates rapid behavioral responses from the cortical network involved in the awareness associated affective response. We showed that the act of immediate rejection of unfair proposals was driven by a phylogenetically old structure (amygdala) and can be viewed as a reactive aggressive response. Moreover, we demonstrated that the amygdala driven rejection response was inhibited with oxazepam treatment without affecting the perception of unfairness. This suggests that the GABA system can influence the decision making network via an alteration of the balance between phylogenetically young (prefrontal cortex) and old structures (amygdala).

As timing is crucial for detection of transient responses (Iidaka et al. 2009) our design had the necessary elaboration to allow detection of fast automatic emotional responses to unfairness and not only slow components. In line with previous studies on emotional bias in decision behavior where the onset time was more precise (De Martino et al. 2006, Paulus et al. 2005) we observed clear amygdala activation. Our study generates two arguments for the causal role of amygdala in the generation of a rejection response. Firstly, in the treatment group, the amygdala response was mitigated in conjunction with a decreased rejection rate. Secondly, in the within subjects comparison in the unmedicated group, rejections were associated with increased amygdala activity. In light of this, we question the suggested causality of insula in the generation of a rejection response that was derived from the correlation between insula activity and acceptance rate of UG offers (Sanfey et al.2003).

A recent study (Knoch et al. 2006) showed that TMS (transcranial magnetic stimulation) of the dlPFC lead to an increased acceptance rate of unfair proposals in the UG without changing the perception of unfairness. The authors concluded that dlPFC drives rejection in response to unfair proposals. That interpretation rests on the notion that the TMS has only local effects with no secondary effects in the neural network underlying decision making (Driver et al.2009). However, that study stands in contrast to previous research. dlPFC has been shown to be necessary for working memory but not for a decision making task involving an emotional bias, while the opposite was true for ventromedial PFC (vmPFC) (Bechara et al.1998). In addition, it has been shown that patients with lesions in vmPFC had an increased rejection rate of unfair proposals in the UG compared to controls (Koenigs and Tranel 2007) in line with the suggestion that this region is involved in regulating emotional conflicts through a direct regulation of amygdala (Etkin et al. 2006). As the prefrontal regions mature over the first years of life the concept of fairness and theory of mind develop across the same age. Playing UG with children would test if dlPFC is necessary for rejection of unfair proposals, i.e. if the behavioral response of rejection rests solely on frontal lobe function. Takagishi and colleagues (Takagishi et al. (2010) have recently demonstrated that preschoolers indeed reject unfair offers in spite of no explicit account of unfairness or theory of mind. Thus, the literature suggests that the vmPFC and not the dlPFC is necessary in decision making biased by an emotional content. In line with this view we observed a relative increase in rACC/vmPFC for unfair vs. fair proposals in the placebo treatment vs. oxazepam treatment, but no changes in the dlPFC. We suggest that this treatment related change is secondary to reduced amygdala input mirroring a reduction of conflict (Etkin et al. 2006, Bush et al. 2000).

We have shown that the basis for decision making in the UG has underpinnings in several brain regions of different phylogenetic origin and this underlines the complexity of responses in the UG. Our data suggest that the automaticity driven rejection response has a phylogenetically older representation as compared to the calculated acceptance based on a consciously determined self-optimizing strategy. The amygdala driven reactive aggressive response generates a behavior that e.g. yields an acceptable splitting of a prey within the group and such an inequity aversion is seen in children (Takagishi et al. (2010) and also in primates (Brosnan and de Waal 2003). Thus, automatic individual reactions to detected unfairness seem to a certain extent support the long term group norms that allow sharing. More developed sharing schemes like formal trade and abstract rule obedience require that each individual can maintain concepts of future effects of present decisions. Such social interactions rest on the development of the human frontal lobe function. We have demonstrated that an anxiolytic drug alters the balance between rapid emotional reactions and reflected feeling based decisions. The finding prompts an ethical discussion as we showed that a commonly used drug influences core functions in the human brain that underlie individual autonomy, economic decision making and thereby social interactions.

4. Methods

4.1. Subjects. Thirty-five right handed volunteers with the mean age of 23.7 $\pm$ 4.2 y (13 men, 22 women) were included in the study. Subjects had no prior or present history of psychiatric illness or neurological disease. All subjects were healthy and took no medications with the exception of birth control pills and mild allergic medications. All participants gave their informed consent. The study was approved by the local ethical committee in Stockholm, Sweden.

4.2. Stimuli/Ultimatum game. Each subject was exposed to 45 different movie clips. In each movie there was a different human proposer who either made a fair, unfair, or neutral suggestion on how to split a sum of money. The fair proposals implied an equal split of the money, saying e.g. “You get 50 Swedish crowns (SEK) and I take 50 SEK”. (7 SEK $\approx$ 1 USD.) The unfair proposals implied that the responder should have 20% and the proposer 80% of the money e.g. “You get 20 SEK and I take 80 SEK”. The total stakes (e.g. 100 SEK) were never mentioned in purpose to maintain ambiguity of fairness until the final proposition of the amount that would be awarded the responder was revealed. All proposals had the exact same wording, except for the total stakes that varied. Subjects were instructed to respond with either yes or no to the proposals. In the neutral control condition the subjects were shown films with proposers saying “this is not a proposal” and subjects were instructed to respond no to these.

The three different stake levels yielded a total of seven different kinds of messages. Each subject encountered six 50/50 offers, seven 20/80 offers, five 125/125 offers, five 50/200 offers, four 250/250 offers, three 100/400 offers, as well as 15 neutral messages. The genders of the proposers were thoroughly balanced (22 males, 23 females).

Before each movie clip the subject was presented with a resting frame containing a hair cross, for a duration that was randomized between 1 and 5 seconds. Thereafter, a film clip with an offer was presented. The onset-times when the proposer finished the sentence were included as regressors of interest (individual regressors for fair, unfair and no proposals, respectively) in the subsequent GLM analysis of the fMRI analysis. Each movie lasted for 7 seconds. The clip was followed by a pause, which was again randomized between 1 and 5 seconds. Thereafter a frame was shown saying "respond now", instructing the subject to make a choice. This frame lasted until a choice had been made or maximally 3 seconds. The onset times when the subject pressed the button were included as a covariate of no-interest in the subsequent GLM analysis of the fMRI analysis. Finally a frame confirmed the decision, saying "you got X SEK, your counterpart got Y SEK", for two seconds.

4.3. Films. A total of 92 persons were filmed and recorded while they read each of the messages as explained above. All persons were filmed under the same conditions; with light from the front on a white background, and with the eyes located in the middle of the screen while speaking into the camera. The films with 45 of these persons were kept and the other were discarded due to low sound quality or because some persons did not look into the camera as desired.

4.4. Monetary reward. Subjects acting as responders were paid 300 SEK for showing up. In addition, three of the 45 movies presented to them were drawn at random, and paid out with real money both for themselves and for the proposer. If the subject had answered yes to such a drawn proposal, both participants were subsequently paid the corresponding amounts of money. In contrast, if the subject had declined the proposal then neither of the two received any money from that film. This information was given to the subjects before the experiment. The persons acting as proposers on the film clips were given 100 SEK for making the films. They were also subsequently paid their part of the proposals that were drawn randomly, as described for the proposers. Average payment to proposers was 380 SEK and average payment to responders was 625 SEK.

4.5. Experimental procedures. Upon arrival, subjects were randomly assigned to either the placebo group (5 men, 12 women) or the oxazepam group (8 men, 10 women). The subjects in the oxazepam group received 20 mg of the drug. Both treatments were administered orally in a single-blind fashion. The subjects had been asked in advance not to eat two hours before the experiment or drink alcohol 24 h prior to the experiment. After drug administration the subjects were asked to fill out two questioners: SSP (Swedish universities Scales of Personality) (Gustavsson et al. 2000) and STAI-t (State Trait Anxiety Index – Trait) (Spielberger et al. 1970). Before entering the scanner the subjects were explained the rules of the UG and their understanding of the game was checked with a questionnaire. All subjects passed this test.

Approximately one hour after treatment the first experimental session was presented. The order in which the film clips were presented was randomized in advance, creating 18 different sequences of clips, or protocols. Each protocol, except for one, was presented for one subject receiving treatment and for one subject in the control group. We used an fMRI-compatible glove "answering device" in the scanner to register the subjects' responses. Subjects responded "yes" by pressing a key with their thumb and "no" by pressing a key with their index finger. All subjects underwent two scanning sessions with a pause of approximately one minute in between. The first session contained 23 movies and the second session contained 22 movies.

After the scanning was completed subjects were asked to rate the fairness of all the kinds of offers they had received, on a scale 1-7 (Knoch et al. 2006). They were also asked to rate the likability of all the faces they had see, on a VAS scale (0-100).

4.6. Statistical analysis.

4.6.1. *Behavioral data.* The effect of the treatment on rejection rate for unfair proposals was first analyzed with a Mann-Whitney U test (one-tailed), since we could not assume normally distributed data. To control for stake size, sex and ordering of decisions we analyzed the individual choices with probit regressions (n.s., see supporting information), since each individual decision is binary in nature. Standard errors were clustered on subjects to account for repeated measures. Since no fair offers were rejected we restricted our attention to the unfair responses. Differences in ratings of fairness and likeability were analyzed with the Mann-Whitney U test. We used two-tailed tests as we had no prior assumption about the direction of a potential treatment effect.

4.6.2. *fMRI data.* All contrasts of interest were initially analyzed on a single-subject level. Four contrasts were of interest when analyzing group data. First, we compared $\text{proposals}_{\text{unfair + fair offers}} > \text{non proposals}_{\text{control condition}}$ in a one sample t-test including all subjects to ensure consistency. As a follow-up we performed a two-sample t-test comparing the two treatment groups in the same contrast. Second, we performed an interaction analysis comparing $\text{placebo}_{\text{unfair – fair proposals}} > \text{oxazepam}_{\text{unfair – fair}}$

proposals. The contrast unfair > fair proposals was also tested within groups to detect the main effect of unfairness. Third, we compared unfair proposals rejected > unfair proposals accepted. In this analysis we included all the subjects (n = 6) in the placebo group that both rejected and accepted unfair proposals in both scanning sessions. Fourth, parameter estimates from right amygdala was analyzed with an ANOVA, specific contrasts were made with a between groups t-test.

4.7. Image acquisition. We used a GE- 1.5 T MR-scanner to measure the blood-oxygen level-dependent (BOLD) responses. A T2* - weighted echoplanar image (EPI) sequence was applied. The following protocol was used: number of slices: 32, slice thickness: 4.5 mm, interslice gap: 0.5 mm, field of view (FOV): 220x220 mm, time echo (TE): 40 ms and time repetition (TR): 2.5 s. 168 and 161 image volumes were acquired during the two scanning sessions, respectively. In addition, we acquired an anatomical T1-weighted 3D image volume from each subject (3D-SPGR, TR/TE = 35 / 6 ms, flip = 35 deg, 124 coronal images, matrix size (0.9 x 1.0 x 0.9 mm3).

4.8. Image analysis.

4.8.1. *Pre-processing.* The functional MRI data were analyzed with the SPM5 software (http://www.fil.ion.ucl.ac.uk/spm/software/). The following pre-processing steps were performed: realignment, slice timing correction, co-registration and normalization with respect to the MNI compatible EPI template provided in SPM5. Finally, spatial smoothing was performed with a Gaussian kernel of 8 mm full-with-half-maximum (FWHM). Event onset-times pertaining to the proposals and control conditions were convolved the canonical hemodynamic response function as implemented in SPM5 and inserted into a general linear model (GLM). Ten regressors were created for each scanning session: (1) unfair proposal, (2) fair proposal, (3) no proposal (control condition) and, (4) reaction time. We corrected for residual movement-related variance in the data by including six motion parameters in the model. High-pass filtering (cut-off frequency = 128 seconds) was used to remove low-frequency noise.

4.8.2. *Regions of interest.* All masks used in the second-level analyses were created with the wfu_pickatlas (Maldjian et al. 2003, Maldjian et al. 2004) tool in SPM5. To validate our study design (timing) we made a global search in the contrast proposals unfair + fair offers > non proposals control condition. We included the following 14 regions of interests (ROI) (bilaterally): anterior cingulate cortex, insula, amygdala, caudate, putamen, medial orbitofrontal cortex, inferior orbitofrontal cortex, superior orbitofrontal cortex, rectus, superior medial frontal cortex, superior frontal cortex, medial frontal cortex, inferior frontal operculum and inferior frontal triangularis. To answer our primary hypothesis we used a bilateral amygdala mask in the contrast placebo $_{\text{unfair –fair proposals}}$ > oxazepam $_{\text{unfair – fair proposals}}$. The same mask was used in the contrast unfair proposals $_{\text{rejected}}$ > unfair proposals $_{\text{accepted}}$. In the extended analyses, the above stated regions (excluding amygdala) were used for global search in the contrasts placebo $_{\text{unfair –fair proposals}}$ > oxazepam $_{\text{unfair – fair proposals}}$ and unfair proposals $_{\text{rejected}}$ > unfair proposals $_{\text{accepted}}$.

4.8.3. *Reporting results.* The SPM [T] map threshold was determined to $p < 0.005$ (uncorrected) in all contrasts. Both corrected and uncorrected results are reported and, hence specified for each peak voxel in the text. All results are reported as voxel level corrected, unless otherwise stated (i.e. cluster level corrected).

Supporting Information

S1. Main effect of proposals. To test the validity of the study design we used the contrast proposals unfair + fair offers > non proposals control condition within all subjects as an overall check of the subject's attention. The analysis showed strong activations bilaterally in the frontal attention network with a peak activation in the right medial prefrontal cortex (mPFC) (MNI space coordinates (x,y,z): [6 24 48] Z score = 6.48; $p < 0.001$ corrected) (Fig. S1A), providing strong support that our subjects indeed actively participated in the experiment. Moreover, we were interested in excluding the possibility that the oxazepam treatment had a general effect on neural activity. Therefore, we tested the interaction placebo unfair+fair proposals – control condition > oxazepam unfair+fair proposals – control condition. This contrast only showed a subsignificant activity in the left supplementary motor cortex (BA 6) (peak voxel: [-24 -15 57] Z score = 4.18, p = 0.11 corrected) (Fig. S1B) possibly reflecting the preparation to make an active choice. In summary, we could conclude that the effect seen in Figure S1A was driven by both groups and not by the placebo group alone. Thus, the general cerebral response in the decision making task was unaltered by the drug.

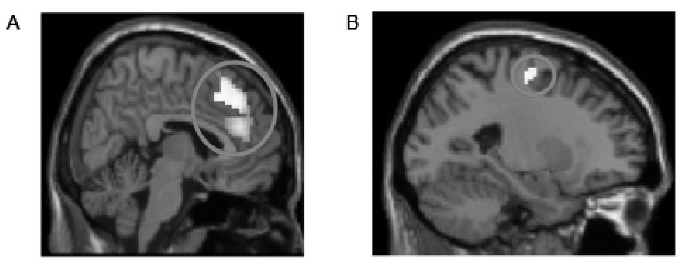

FIGURE S1. *Neural activity related to receiving a proposal.* (A) As a manipulation check we compared proposals unfair + fair offers > non proposals control condition within all subjects. The contrast showed an activation in the frontal attention network with a peak activation in the right mPFC (MNI space coordinates (x,y,z): [6 24 48]; Z score = 6.48; $p < 0.001$ corrected). (B) The interaction contrast treatment (placebo > oxazepam) × proposal (unfair + fair > control condition) showed a subsignificant activation in the left supplementary motor cortex (BA 6) ([-24 -15 57] Z score = 4.18, p = 0.11 corrected).

S2. Behavioral data for rejection rate.

S2.1. *Robustness to stake effects.* Stake levels did not influence rejection rate as the effect of treatment on rejection rate remained after controlling for the different stake levels, in a probit regression (t-test, one-tailed, Z=1.67, p=0.0475).

S2.2. *Robustness over time.* To test if rejection rate changed over time we conducted a probit regression where we included the order of the decision (1-45) as an explanatory variable, together with dummy variables for the stake levels. The order variable did not contribute to explaining decisions to reject offers. The coefficient on the order variable was -0.0038548 and insignificant (t-test, two-tailed, Z=-0.99, p=0.323). A similar result was observed when a dummy variable for the second half of the films was used as an explanatory variable (t-test, two-tailed, Z=-0.48, p=0.633).

S2.3. *Sex differences.* The effect of treatment did not reveal any significant sex differences. Women's average rejection rate dropped by 15% from 40% to 25% by treatment (Mann-Whitney U test, two-tailed, Z=1.116, p=0.2645). Men's average rejection rate dropped by 20% as an effect of treatment, from 31% to 11% (Mann-Whitney U test, two-tailed, Z=1.036, p=0.3002). The differences in drops were not significant in probit regression analysis when rejection rate was regressed on sex, treatment, and sex *treatment. Thus, the point estimate of the interaction coefficient was insignificant (t-test, two-tailed, Z=0.20, p=0.846).

S3. The insula. Although we did not observe any significant activation in insula for unfair vs. fair proposals we present our data on the insula since it has been suggested that this structure is associated to rejection of unfair proposals (Sanfey et al.2003). The main effect of unfairness (unfair – fair proposals) in the oxazepam group showed a subsignificant activation in right insula ([36 21 12] $Z = 3.34$, $p < 0.001$ uncorrected) (Fig. S2). The same contrast in the placebo group showed a subsignificant activation in the left insula ([-30 24 3] $Z = 2.90$, $p < 0.005$ uncorrected). These Z-scores are in the same range as the significance values presented for insula activation in a previous imaging study of the ultimatum game (Sanfey et al.2003) however, we implemented a more conservative statistical correction procedure. Importantly, we did not observe any difference between the two treatment groups in processing unfair proposals (placebo unfair – fair proposals > oxazepam unfair - fair proposals). Thus, the observed change in choice behavior between the treatment groups was not explained by an altered insula activity.

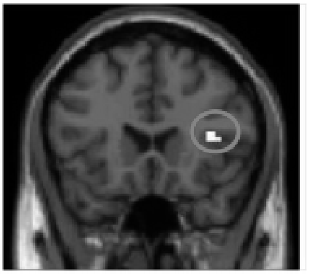

FIGURE S2. Figure S2. *fMRI results related to unfair proposals.* In the oxazepam group a subsignificant activation was present in the right insula ([36 21 12] Z score = 3.34, $p < 0.001$, uncorrected).

References

Tversky A, Kahneman D (1981), 'The framing of decisions and the psychology of choice', *Science* 30, 453-458.

De Martino B, Kumaran D, Seymour B, Dolan RJ (2006), 'Frames, biases, and rational decision-making in the human brain', *Science* 313, 684-687.

Camerer C (2003), *Behavioral game theory: Experiments in strategic interaction*, Princeton University Press, Princeton, NJ.

Bechara A, Damasio H, Damasio A (2003), 'Role of the amygdala in decision-making', *Annals of the New York Academy of Sciences* 985, 356-369.

Craig AD (2009), 'How do you feel now? The anterior insula and human awareness', *Nat Rev Neurosci* 10, 59-70.

Pinker S (2003), *The blank slate: The modern denial of human nature*, Penguin Books.

Fehr E, Gächter S (2000), 'Cooperation and punishment in public goods experiments', *American Economic Review* 90. 980-994.

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003), 'The neural basis of economic decision-making in the ultimatum game', *Science* 300, 1755-1758.

Tricomi E, Rangel A, Camerer CF, O'Doherty J (2010), 'Neural evidence for inequality-averse social preferences', *Nature* 463, 1089-1091

Buckholtz JW, Asplund CL, Dux PE, Zald DH, Gore JC, Jones OD, Marois R (2008), 'The neural correlates of third-party punishment', *Neuron* 60, 930-940.

Güth W, Schmittberger R, Schwarze B (1982), 'An experimental analysis of ultimatum bargaining', *Journal of Economic Behavior and Organization* 3, 367-388.

Wallace B, Cesarini D, Lichtenstein P, Johannesson M (2007), 'Heritability of ultimatum game responder behavior', *Proceedings of the National Academy of Sciences* 104, 15631-15634.

Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, et al. (2005), '"Economic man" in crosscultural perspective: behavioral experiments in 15 small-scale societies' *Behavioral and Brain Science* 28, 795–815.

Vuilleumier P (2005), 'How brains be aware' *Trends in Cognitive Science* 9, 585-594.

Bechara A, Damasio H, Tranel D, Anderson SW (1998), 'Dissociation of working memory from decision making within the human prefrontal cortex', *J. Neurosci.* 18, 428-437.

Xue G, Lu Z, Levin I, Bechara A (2010), 'The impact of prior risk experiences on subsequent risky decision-making: The role of the insula', *NeuroImage* 50, 709-716.

Damasio AR (1994), *Decartes' error - emotion, reason, and the human brain*, Grosset/Putnam, New York.

Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, Cohen JD (2004), 'The neural correlates of theory of mind within interpersonal interactions', *NeuroImage* 22, 1694-1703.

Vuilleumier P, Armony JL, Driver J, Dolan RJ (2001), 'Effects of attention and emotion on face processing in the human brain: An event-related fmri study', *Neuron* 30, 829-841.

Nelson RJ, Trainor BC (2007), 'Neural mechanisms of aggression', *Nat Rev Neurosci* 8, 536-546.

Arce E, Miller D, Feinstein J, Stein M, Paulus M (2006), 'Lorazepam dose-dependently decreases risk-taking related activation in limbic areas', *Psychopharmacology* 189, 105-116.

Paulus MP, Feinstein JS, Castillo G, Simmons AN, Stein MB (2005), 'Dose-dependent decrease of activation in bilateral amygdala and insula by lorazepam during emotion processing', *Arch Gen Psychiatry* 62, 282 - 288.

Knoch D, Pascual-Leone A, Meyer K, Treyer V, Fehr E (2006), 'Diminishing reciprocal fairness by disrupting the right prefrontal cortex', *Science* 314, 829-832.

Iidaka T, Saito D, Komeda H, Mano Y, Kanayama N, Osumi T, Ozaki N, Sadato N (2009), 'Transient neural activation in human amygdala involved in aversive conditioning of face and voice', *J Cogn Neurosci.*, 2074-2085.

Driver J, Blankenburg F, Bestmann S, Vanduffel W, Ruff CC (2009), 'Concurrent brain-stimulation and neuroimaging for studies of cognition', *Trends in Cognitive Science* 13, 319-327.

Koenigs M, Tranel D (2007), 'Irrational economic decision-making after ventromedial prefrontal damage: Evidence from the ultimatum game', *J. Neurosci.* 27, 951-956.

Etkin A, Egner T, Peraza DM, Kandel ER, Hirsch J (2006), 'Resolving emotional conflict: A role for the rostral anterior cingulate cortex in modulating activity in the amygdala', *Neuron* 51, 871-882.

Takagishi H, Kameshima S, Schug J, Koizumi M, Yamagishi T (2010), 'Theory of mind enhances preference for fairness', *Journal of Experimental Child Psychology* 105, 130 -137.

Bush G, Luu P, Posner MI (2000), 'Cognitive and emotional influences in anterior cingulate cortex', *Trends in Cognitive Science* 4, 215-222.

Brosnan SF, de Waal FBM (2003), 'Monkeys reject unequal pay', *Nature* 425, 297-299.

Gustavsson PJ, Bergman H, Edman G, Ekselius L, von Knorring L, Linder J (2000), 'Swedish universities scales of personality (ssp): Construction, internal consistency and normative data', *Acta Psychiatrica Scandinavica* 102, 217-225.

Spielberger C, Gorsuch R, Lushene R (1970), *Manual for the state-trait anxiety inventory (self evaluation questionnaire)*, Consulting Psychologists Press Palo Alto, CA.

Maldjian J, Laurienti P, Burdette J, Kraft R (2003), 'An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fmri data sets', *NeuroImage* 19, 1233-1239.

Maldjian J, Laurienti P, Burdette J (2004), 'Precentral gyrus discrepancy in electronic versions of the talairach atlas', *NeuroImage* 21, 450-455.

PAPER 3

Evolution of Theories of Mind

Erik Mohlin

Abstract. This paper studies evolution of peoples' models of how other people think – their theories of mind. For the case of games that are played for the first time, people are assumed to form beliefs according to the level-k model. This model postulates a hierarchy of types, such that an individual of type k plays a k times iterated best response to the uniform distribution. For the case of learning, it is assumed that the lowest type behaves in accordance with fictitious play, and that there is a hierarchy of more sophisticated types, which play iterated best responses to this. The models are also extended to allow for partial observability, in the sense that a higher type recognize and best respond to lower types, but not vice versa. Evolution according to the replicator dynamic is studied both across and within games. It is found that evolution may lead to stable states where different types, including low types, co-exist. This holds even when types are not observed.

1. Introduction

In order to decide what strategy to choose, a player needs to form beliefs about what other players will do. This requires the player to have a model of how other people form beliefs – what psychologists call a *theory of mind* (Premack and Wodruff 1979). In this paper I study the evolution of theories of mind, both in the form of models of how other players form initial beliefs, and in the form of models of how other people learn.

When people play a game for the first time, their behavior rarely conforms to a Nash equilibrium.[1] In such situations, behavior is more successfully predicted by the *level-k*

Valuable comments were provided by Ola Andersson, Vincent Crawford, Tore Ellingsen, Drew Fudenberg, and Robert Östling. A special thank goes to Jörgen Weibull for advice and encouragement. The paper has also benefitted from comments made by audiences at the European Winter Meeting of the Econometric Society, Budapest 2009, the 1st UECE Lisbon Meeting: Game Theory and Applications 2009, the 4th Nordic Workshop in Behavioral and Experimental Economics, Oslo 2009, as well as the University of Bonn (Neureconomics Lab), University of Amsterdam (CREED), University College London, and the Stockholm School of Economics. Financial support from the Jan Wallander and Tom Hedelius Foundation is gratefully acknowledged.

[1] See Goeree and Holt (2001) and Camerer (2003). In order to claim that Nash equilibrium predictions are refuted by behavioral data one must make some assumption about preferences. In the mentioned studies, the preferences needed to make the observed behavior conform to Nash equilibrium

(Stahl and Wilson 1995 and Nagel 1995) and *cognitive hierarchy* models (Camerer et al. 2004), or models of noisy introspection (Goeree and Holt 2004).[2] According to these models, people think in a limited number of steps, when they form beliefs about other peoples' behavior. Moreover, people differ with respect to how they form beliefs. The heterogeneity is represented by a set of cognitive types $\{0, 1, 2, ...\}$, such that higher types form more sophisticated beliefs. Type 0 does not form any beliefs and randomizes uniformly over the strategy space. According to the level-k model, an individual of type $k \geq 1$ believes that everyone else belongs to type $k - 1$. All types $k \geq 1$ best respond given their beliefs, and have identical preferences.[3] Empirically one finds that most experimental subjects behave as if they are of type 1 or 2, and individuals of type 3 and above are very rare (Costa-Gomez and Crawford 2006, Camerer 2003).[4]

In order to study evolution of theories of mind in the context of learning, I consider an extension of fictitious play. According to fictitious play all individuals believe that the future will be like the past, and best respond to the average of past play. If some individuals follow this rule, it is natural to hypothesize that some more sophisticated individuals could understand that other individuals behave in accordance with fictitious play. These more sophisticated individuals would then play a best response to the best response to the average of past play. And it seems quite possible that some individuals think yet another step and play a twice iterated best response to the average of past play. Continuing in this way we arrive at a hierarchy of types $\{1, 2, ...\}$, using increasingly complex models of how other people learn. I refer to the resulting model as *heterogeneous fictitious play.* A related model is proposed, and tested, by Stahl (1999, 2000).[5]

---

seem like a much less reasonable explanation than attributing the behavior to some form of incorrect expectations. See the discussion in (**?**).

[2] For experimental evidence on this see Camerer et al. (2004), Costa-Gomez and Crawford (2006), and Camerer (2003). Coricelli and Nagel (2009) present neuroeconomic evidence. The models have also been applied to e.g. auctions (Crawford and Iriberri 2007), communication (Crawford 2003, Kawagoe and Takizawa 2008, Ellingsen and Östling 2009), and marketing decisions (Brown et al. 2008).

[3] In the cognitive hierarchy model, an individual of type $k \geq 1$ believes that everyone else belong to type 0 through $k - 1$, and has a correct belief about the relative population fractions of these lower types. The noisy introspection model is similar to the level-k model except for the fact that all types play a noisy best repy and are aware of the fact that other types also play a noisy best response.

[4] This estimate holds regardless of whether types are defined according to the level-k or the cognitive hierarchy model. The results are also robust to tests of the presence of many alternative types, such as a type that always plays the Nash equilibrium, or a type that play a best response to the actual distribution of play. It is usually found that type 0 is quite rare, indicating that it should perhaps be interpreted as existing mostly in the minds of higher types (see e.g. Costa-Gomez and Crawford 2006).

[5] For discussions of the empirical support of different learning models see Salmon (2001), Camerer (2003), and Wilcox (2006). Note that Wilcox's criticism does not seem to apply to models that

Many games with important consequences are only played a few times during a life time time, with little scope for learning. For example, the choice of a career and the choice of a mate are parts of complicated games which most people play only once, or a few times. Still, the strategic thinking employed in such interactions should be highly relevant for the success of a person, in economic as well as biological terms. Thus, strong evolutionary forces work in favor of those who do well the first time they play games. Other games are played many times, with feedback that allows the players to learn. The evolutionary advantage of an accurate model of how other people learn should be obvious in such cases. The main contribution of this paper is to provide an evolutionary analysis of the level-k, cognitive hierarchy, and heterogeneous fictitious play models. In particular I explain why evolution may lead to a state where people display heterogeneous and limited strategic thinking, often resulting in non-Nash behavior initially.[6] There are only a few studies of evolution of cognitive types which can be interpreted as being about initial responses; Stahl (1993), Banerjee and Weibull (1995), and Stennek (2000). These models differ in important aspects from the model put forward in this paper, and they do not identify the same mechanisms forming the distribution of types (see section 5). There is a vast literature studying properties of different formal learning rules (see e.g. Weibull 1995, Fudenberg and Levine 1998, and Sandholm 2010), but there has hardly been any studies of the evolutionary properties of the learning rules themselves (an exception is Josephson 2008).

The level-k and cognitive hierarchy models, as well as fictitious play, implicitly assume that players lack specific information about the cognitive types of their opponents. A second contribution of this paper is to extend these models to allow for the possibility that types are partially observed. Such an extension is essential in order to capture situations where an unfamiliar game is played by individuals who have some information about their opponents' ways of forming beliefs. Such information may stem from previous interactions in other games, or from communication. In the kind of small-scale societies that characterized much of our evolutionary past, such information was probably common. To model this I assume that higher types can recognize and best respond to lower types, but that lower types are unable to (fully) understand how the higher types think. These assumptions follow naturally from the way that types are defined. The cognitive type of an individual represents that individual's ability to

postulate heterogeneity, like the model of Stahl, or the model suggested here. Note also that Salmon explicitly says that his analysis does not apply to Stahl's models.

[6] One might object that the heterogeneity may be due to random variation. However, there is evidence that strategic reasoning is implemented by specialized modules in the brain (Cosmides and Tooby 1992), and it has been argued (Penke et al. 2007) that variation in such traits is best explained by frequency dependent selection, rather than with random variation.

understand how other people think. Thus, being of a high type means that one is good at understanding how other people think, i.e. that one is good at detecting what type they are.[7]

For the evolutionary analysis I consider a large population of individuals of different types, who are randomly matched to play a symmetric two-player game. When analyzing heterogeneous fictitious play I assume that the drawn individuals play the same game many times. The types of the matched pair of individuals, together with the composition of types in the population, determine the individuals' beliefs, and hence their actions and payoffs, in the game. The population fractions of types evolve in proportion to the average payoffs of different types. Formally, this is done by applying the replicator dynamic to the type space. Most of the literature on evolution and learning in games focuses on one game at a time.[8] In addition to studying evolution in fixed games, I also study the evolution of types across different games. This is done by assuming that individuals are randomly matched to play games that are drawn from a class of games.

From an evolutionary perspective, the potential advantage of a better theory of mind has to be weighted against the cost of increased reasoning capacity. I abstract from such costs in the formal analysis, but note that they limit the survival chances of higher types. Thus, by excluding cognitive costs from the formal analysis I "tilt the board against myself".

I restrict attention to symmetric two-player games. The analyzed games include all *2*-strategy games; coordination games such as the Stag Hunt game, games with a unique interior evolutionarily stable strategy (ESS), like the Hawk Dove game, and dominance solvable games, such as the Prisoners' Dilemma. Most results can be generalized to n-strategy games. The n-strategy coordination games are then defined by the property that each pure strategy is the unique best reply to itself. The *2*-strategy games with a unique interior ESS are generalized to n-strategy cyclic games, or stable cyclic games. The *2*-strategy dominance solvable games are generalized to games defined by a property that I call 'weak best reply dominance'. This class of games includes strictly supermodular games. It also includes e.g. the Travelers' Dilemma (Basu 1994). Like the Prisoners' Dilemma, this is a social dilemma, in which all players would be

[7] In the literature on preference evolution it is well known, that if players have complete information about preferences, then evolution may lead to states where people have preferences that do not coincide with material payoffs (Dekel et al. 2006). As pointed out e.g. by Samuelson (2000b), it is important that assumptions about observability are convincingly motivated and not ad hoc.

[8] A notable exception is Haruvy and Stahl (2009) who experimentally investigate learning across games with different strategy spaces. I know of no theoretical investigation of evolution or learning across games with different strategy spaces. Mengel (2009) models evolution, and Steiner and Stewart (2008) learning, across games with identical strategy spaces.

better of if they could commit to a cooperative strategy, instead of playing the unique Nash equilibrium.

First consider the case of unobserved types. In all of the mentioned dominance solvable games, evolution wipes out all types that are not sophisticated enough to be able to form beliefs that induce them to play a Nash equilibrium. However, for cyclic games (e.g. Hawk Dove games), I find that all states are unstable, in which some type, except type 0, is absent. In some cyclic games, such as the Hawk Dove game, there is a globally attracting asymptotically stable set of states which includes states where all types co-exist, and does not include any state where only one type exists. The intuition for this result is that there is an advantage of not thinking like the opponent, since choosing the same action as the opponent yields a low payoff to both players. In the game of Shapley (1964), I find that evolution in the heterogeneous fictitious play model converges to a state where different types co-exist so that behavior corresponds to the Nash equilibrium in all periods. In coordination games everyone except type 0 earns the same and survives. Finally I establish conditions that assure that a heterogeneous population is preserved in the case of evolution across a mix of these games.

When partial observability is allowed, the survival prospects of lower types are strengthened further. In the Travelers' Dilemma there might now be asymptotically stable states in which some lower types survive. The intuition for this result is that if types are observed, then it is as if lower types are committed to strategies that result in higher payoffs for both players. In the cyclic games introduction of partial observability may lead to more or less heterogeneity, depending on the details of the payoffs. In Hawk-Dove games, evolution from any interior initial state converges to a unique interior state where all types, except type 0, co-exist. Again it can be show that evolution across a mix of these games, may result in heterogeneity.

In the main model all results are derived for the level-k model. All results carry over to the model of heterogeneous fictitious play and the cognitive hierarchy model. Additionally, for the cognitive hierarchy model, I identify conditions under which there is an asymptotically stable fraction of type 0. I also discuss an alternative specification for the case of partially observed types, and conclude that the results are robust. Furthermore, I show that the main conclusions are robust to the introduction of a *Nash equilibrium type*, which is preprogrammed to a Nash equilibrium strategy. The intuition for this is that it may be unprofitable to be a hyper-sophisticated Nash player when other individuals are unsophisticated.

The rest of the paper is organized as follows: The next section presents the basic model; the evolutionary set-up, the cognitive types according to the level-k model, and the underlying games. The results for the level-k model are presented in section

3. Section 4 discusses the heterogeneous fictitious play mode, the cognitive hierarchy model, the Nash type, and the alternative specification of behavior for the case of partially observed types. Section 5 contains some discussion of the results and related literature. Section 6 concludes. All proofs are in the appendix.

2. Model

This section covers the main model. In order to make the exposition accessible it is only concerned with initial play by cognitive types defined according to the level-k model. In section 4, I extend the framework to a model of learning, heterogeneous fictitious play. I also extend the framework to others models of initial play, including the cognitive hierarchy model, and a Nash equilibrium type.

2.1. Preliminaries . Consider a symmetric two-player normal form game G with a finite pure strategy set S and mixed strategy set $\Delta(S)$. Payoffs are given by $\pi : S \times S \to \mathbb{R}$, where $\pi(s, s')$ is the payoff to a player using strategy s against strategy s'. For mixed strategies the expected payoffs are given by $\tilde{\pi} : \Delta(S) \times \Delta(S) \to \mathbb{R}$ where $\tilde{\pi}(\sigma, \sigma')$ is the payoff to player, using strategy σ against strategy σ'. With slight abuse of notation let s denote the degenerate mixed strategy that puts all weight on pure strategy s. Thus $\tilde{\pi}(s, s')$ stands for the expected payoff to a player using the mixed strategy that put all weight on the pure strategies s against the mixed strategy that put all weight on the pure strategies s'. Let $\beta : \Delta(S) \twoheadrightarrow S$ be the *pure best reply correspondence*. If the best response is unique I write $\beta(\sigma) = s$ rather than $\beta(\sigma) = \{s\}$. The *mixed best reply correspondence* is $\tilde{\beta} : \Delta(S) \twoheadrightarrow \Delta(S)$. The uniform randomization over the set of pure best responses to σ, is denoted $\bar{\beta}(\sigma)$. Again, with slight abuse of notation, the expression $\beta(s)$ stands for the pure best response to the mixed strategy that puts all weight on the pure strategy s (and similarly for $\tilde{\beta}$ and $\bar{\beta}$).

Consider a population consisting of a finite set of cognitive types $K = \{0, 1, 2, ..., \kappa\}$. The set of probability distributions over K is $\Delta(K)$, so a *population state* is a point

$$x = (x_0, x_1, ...x_\kappa) \in \Delta(K).$$

Suppose that two individuals from this population play a symmetric two-player normal form game G. All individuals of the same type k play the same strategy $\sigma(k, k') \in \Delta(S)$, against individuals of type k'. The weight put on pure strategy s is $\sigma_s(k, k')$.[9] Let $\sigma(K, x)$denote aggregate play at state x.

[9] In the standard level-k model behavior of type k does not depend on the opponent's type k', but in my model of partially observed types behavior will depend on k'. Later, when considering the model of heterogeneous fictitious play, as well as the cognitive hierarchy model, behavior will also depend on the state.

For a given game G, the expected payoff of type k, against type k', is

$$\Pi_k^G(k') = \sum_{s\in S}\sum_{s'\in S}\sigma_s(k,k')\,\sigma_{s'}(k',k)\,\pi(s,s').$$

Thus, the expected payoff of type k, in state x, is given by the function $\Pi_k^G : \Delta(K) \to \mathbb{R}$, with

$$\Pi_k^G(x) = \sum_{k'\in K} x_{k'}\Pi_k^G(k').$$

Suppose (in order to examine evolution across games) that individuals are randomly matched to play a game which is drawn from a finite set of games $\mathcal{G}$, according to a probability measure μ. The expected payoff of type k, in state x, is now given by the function $\Pi_k^{\mathcal{G}} : \Delta(K) \to \mathbb{R}$, with

$$\Pi_k^{\mathcal{G}}(x) = \sum_{G\in\mathcal{G}} \mu^G\Pi_k^G(x),$$

where μ^G is the probability of game G.

2.2. Evolution . For a given game G, the average payoff in the population, in state x, is

$$\bar{\Pi}^G(x) = \sum_{k=0}^{\kappa} x_k\Pi_k^G(x).$$

Evolution of types is determined by the *replicator dynamic*

$$\dot{x}_k = [\Pi_k^G(x) - \bar{\Pi}^G(x)]x_k.$$

Similarly if the games are drawn from $\mathcal{G}$ according to μ, the average payoff in the population, in state x, is denoted $\bar{\Pi}^{\mathcal{G}}(x)$ and the replicator is defined as above, with $\mathcal{G}$ instead of G.

In the level-k model, each type's behavior is constant across states. It is then easy to verify that the payoffs to the different types, and hence the vector field, is Lipchitz continuous. By the Picard-Lindelöf theorem the system therefore has a unique solution $\xi(\cdot,x^0) : T \to \Delta(K)$ through any initial condition x^0, such that $\xi(0,x^0) = x^0$ and

$$\frac{\partial}{\partial t}\left(\xi\left(t,x^0\right)\right) = [\Pi^G\left(\xi\left(t,x^0\right)\right) - \bar{\Pi}^G\left(\xi\left(t,x^0\right)\right)]\xi\left(t,x^0\right),$$

for all $x \in \Delta(K)$.[10]

[10] In the model of heterogeneous fictitious play, and in the cognitive hierarchy model, behavior will generally not be continuous across states. As a consequence, the vector field will generally not be Lipchitz continuous in the state. However, the vector field will be Lipchitz continuous almost everywhere. The reason is that behavior only changes in states where some type is indifferent between two or more strategies. This set is constituted by the union of a finite set of hyperplanes in the type space. These hyperplanes divide the type space into a finite number of open set. Within each of these

We are interested in Lyapunov stable and asymptotically stable states of the replicator dynamic. For reasons that will become clear below, asymptotically stable sets are also of importance. Hence we need the following definitions:[11]

DEFINITION 1. *A closed set $A \subset \Delta(K)$ is* ***Lyapunov stable*** *if every neighborhood B of A contains a neighborhood B^0 of A, such that if the system starts in $B^0 \cap \Delta(K)$ at time t_0, then the system remains in B at all times $t \geq t_0$.*

A closed set $A \subset \Delta(K)$ is ***asymptotically stable*** *if it is Lyapunov stable and if there exists a neighborhood B^* of A such that if the system starts in B^* at t_0 then as $t \to +\infty$ the system goes asymptotically to A.*

The ***basin of attraction*** *of a closed set $A \subset \Delta(K)$ is the set of states such that starting from such a state the system goes to A as $t \to +\infty$.*

A set $A \subset \Delta(K)$ is an ***attractor*** *if its basin of attraction is a neighborhood of A.*

Stability of a point is defined as the stability of the singleton $\{x\}$. Note that a Lyapunov stable set is asymptotically stable if and only if it is an attractor. A state is *polymorphic* if it contains positive fractions of more than one type. Otherwise the state is *monomorphic*. Finally the concept of an evolutionarily stable strategy (ESS) will be used:

DEFINITION 2. *A strategy $\sigma \in \Delta(S)$ is an* ***evolutionarily stable strategy (ESS)*** *if (i) $\tilde{\pi}(\sigma', \sigma) \leq \tilde{\pi}(\sigma, \sigma)$ for all $\sigma' \in \Delta(S)$, and (ii) $\tilde{\pi}(\sigma', \sigma) = \tilde{\pi}(\sigma, \sigma)$ implies $\tilde{\pi}(\sigma', \sigma') < \tilde{\pi}(\sigma, \sigma')$ for all $\sigma' \neq \sigma$.*

For more on these concepts, and their relations, see Weibull (1995).

2.3. The Level-*k* Model .

2.3.1. *Unobserved Types.* According to the *level-k* (henceforth *LK*) model, type 0 randomizes uniformly over the strategy space. All types $k \geq 1$ best reply given their beliefs. If the pure best reply is not unique, then the individual is assumed to follow the principle of insufficient reason and randomize uniformly over the set of pure best replies. Let U denote the uniform distribution over S, and let $\beta^i(U)$ denote the i times iterated best response to the uniform distribution. Recall that $\bar{\beta}$ denotes the uniform randomization over the set of pure best replies. Thus the behavior of type $k \geq 1$ is

$$\sigma(k) = \bar{\beta}^k(U).$$

Note that behavior is independent of the opponent's behavior.

sets, behavior is constant across states. This allows us to use a definition of a solution in Filippov's (1960) sense.

[11] By a *neighborhood* of a closed set A is meant an open set B such that $A \subseteq B$.

2.3.2. *Partially Observed Types .* As explained in the introduction, there are many situations in which people play an unfamiliar game with people they already know something about. In particular players may have information about their opponents' theories of mind. The LK model assumes that individuals do not observe each other's type. In this section I propose one simple way of relaxing this assumption and extending the LK model to the case of partially observed types. The only modification that I add to the LK model is that when an individual of type k faces an opponent of a lower type $k' < k$ then the former is able to understand how the latter thinks, and hence the higher type best responds to the lower type. The lower type is assumed to behave exactly as in the ordinary level-k model with unobserved types. Formally;

$$\sigma(k, k') = \begin{cases} \bar{\beta}(U) & \text{against } k' = 0 \\ \bar{\beta}^{k'+1}(U) & \text{against } k' \in \{1, 2, ...k-1\} \\ \bar{\beta}^{k}(U) & \text{against } k' \geq k \end{cases}.$$

In section 4.4, I discuss an alternative specification and conclude that it yields qualitatively similar results.

2.4. The Underlying Games . In this section I define the games that are included in the main analysis. I only study symmetric *2*-player games, and throughout the paper I assume that the Nash equilibrium is different from the uniform distribution; $U \neq \sigma^{NE}$. All generic and symmetric 2×2 games fall into one of three categories of strategically equivalent games.[12] Suppose $a, b > 1$. *Coordination games* have payoffs of the form

$$\begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix}.$$

Such games have two symmetric pure strategy equilibria, both of which correspond to evolutionarily stable strategies (ESS), and one symmetric mixed strategy equilibrium, which does not correspond to an ESS. The Stag Hunt Game falls into this category. More generally define:

Definition 3. *An n-strategy* ***coordination game*** *is such that $\beta(s) = s$ for all $s \in S$.*

Games with a unique interior ESS have payoffs of the form

$$\begin{pmatrix} -b & 0 \\ 0 & -1 \end{pmatrix}.$$

[12] By 'generic' is meant that there are no payoff ties. Two symmetric two-player games are strategically equivalent if they share the same dominance relations and best reply correspondences. Formally, two symmetric two-player games with payoff matrices $\mathbf{A} = \{a_{ij}\}$ and $\mathbf{A}' = \{a'_{ij}\}$ are strategically equivalent if $a_{ij} = \lambda a'_{ij} + \mu + v_j$ for some $\lambda \in \mathbb{R}_{++}$, $\mu \in \mathbb{R}$, and $\nu_j \in \mathbb{R}$.

Such games have two asymmetric pure strategy equilibria and one symmetric mixed strategy equilibrium, where only the latter corresponds to an ESS. The Hawk Dove Game falls into this category. The Hawk Dove game is cyclic in the sense that the best response to strategy s is strategy $s + 1_{\text{mod}\,2}$. I will therefore work with the following generalization:

Definition 4. *An n-strategy* ***cyclic game*** *is such that $\beta(s) = s + 1_{\text{mod}\,n}$ for all $s \in S$.*

Moreover, the Hawk Dove game is stable game. A normal form game with payoff matrix $\mathbf{A}$ is said to be *stable* if $\mathbf{A}$ is negative definite with respect to the tangent space. Formally, this requires that $v \cdot \mathbf{A}v < 0$ for all $v \in \mathbb{R}_0^n = \{\mathbb{R}^n : \sum v_i = 0\}$, $v \neq \mathbf{0}$, where $\mathbb{R}_0^n$ is called the tangent space.

Finally there are *dominance solvable games*, such as the Prisoners' Dilemma

$$\begin{pmatrix} -a & 0 \\ 0 & b \end{pmatrix}.$$

The Prisoners' Dilemma has two properties that I wish to generalize. The first one is that it is dominance solvable. The second one is that the Nash equilibrium is Pareto dominated by another symmetric strategy profile. To this end the *Travelers' Dilemma* (Basu 1994), is included.

Definition 5. *The* ***Travelers' Dilemma*** *is a symmetric two-player normal form game with strategy space $S = \{0, 1, 2, ..., c\}$, for some $c \in \mathbb{N}$. The payoff to a player choosing strategy s against strategy s' is*

$$\pi(s, s') = \begin{cases} s & \text{if } s = s' \\ s + R & \text{if } s < s' \\ s' - P & \text{if } s > s' \end{cases},$$

for some real numbers $R, P > 1$ with $R + P = l + r$, for some $l \in \mathbb{N}$, and some $r \in (0, 1)$.

The assumption that $R + R \notin \mathbb{N}$ is made in order to assure that all types $k \geq 1$ have a unique best response given their beliefs. This assumption only serves to simplify the exposition. In the Travelers' Dilemma there is always an incentive to undercut the opponents choice – by picking a strategy that is one step below the opponent's strategy one obtains a net reward of $R - 1 > 0$. Therefore this game (like the Prisoners' Dilemma) constitutes a social dilemma, in the sense that both players would earn more if they were are able to cooperate and play a high strategy, than if they play the Nash equilibrium $(0, 0)$.

I will also work with a more general class of dominance solvable games.

DEFINITION 6. *A game satisfies* ***weak best reply dominance (WBRD)*** *if*

$$\pi\left(\beta\left(s\right),\beta^{k}\left(s\right)\right)\geq\pi\left(s,\beta^{k}\left(s\right)\right)$$

for all $k \geq 0$ *and all* s.

This property is satisfied by strictly supermodular games. It is also satisfied by the following class of games:

DEFINITION 7. *A game satisfies* ***ordered single-peaked payoffs (OSPP)*** *if the strategy space can be ordered* $S = \{1, 2, ..., n\}$, *such that; for all* $s' < n$, $\pi(\cdot, s')$ *is single peaked with the unique maximum at some strategy* $s > s'$, *and* $\pi(\cdot, n)$ *is single-peaked with the unique maximum at* n.

The relationship between weak best reply dominance, ordered single-peaked payoffs, and strict supermodularity is described by the following lemma:

LEMMA 1. ***(a)*** *If a game satisfies OSPP or is strictly supermodular, then it satisfies WBRD.* ***(b)*** *Strict supermodularity does not imply OSPP, and OSPP does not imply supermodularity.*

A game that is not supermodular, but satisfies the OSPP property is the Travelers' Dilemma.

When the type space is restricted to $K\backslash\{0\}$, i.e. when type 0 is excluded, I will be able to prove results for the more general n-strategy games. When type 0 is included it will often be necessary to restrict attention to the less general classes of games (*2*-strategy games and the Travelers' Dilemma) in order to obtain results. Furthermore, in order to obtain results for the case of partially observed types I will often have to restrict attention to the less general classes of games.

3. Results for Initial Play: Level-k Types

3.1. Unobserved Types. First consider coordination games. Since each pure strategy is the unique best response to itself, all types $k \geq 1$ behave in the same way. Consequently all types $k \geq 1$ earn the same in all states, and all types $k \geq 1$ earn more than type 0. We arrive at the following simple result.

PROPOSITION 1. *Suppose that an underlying coordination game is played by unobserved LK types. Evolution from any interior initial state converges to some state where* $x_0 = 0$ *and where all other types exist in the same relative fractions as in the initial state. The set of all such states is the unique asymptotically stable set.*

Thus in coordination games there is no evolutionary advantage of belonging to a high type, as long as one does not belong to type 0. Games satisfying weak best reply dominance (WBRD), such as the Travelers' Dilemma, are a less friendly environment for low types. We find that higher types have a strict advantage over lower types. Let S^{NE} denote the set of pure strategy Nash equilibria. Let $\tilde{k}$ be the minimum number of iterated best responses to the uniform distribution, which is required to reach a Nash equilibrium. Formally,

$$\tilde{k} = \min\left\{i \in \mathbb{N} : \beta^{i}(U) \in S^{NE}\right\}.$$

All types $k \geq \tilde{k}$ will behave in exactly the same way. Therefore, in what follows I will assume that $\kappa \leq \tilde{k}$, or equivalently $\beta^{\kappa-1} \notin S^{NE}$. In other words, the strategy space of WBRD-games is assumed to be sufficiently "rich" relative to the type space, so that all types distinguish themselves behaviorally.

Proposition 2. ***(a)*** *Suppose that an underlying game, which satisfies WBRD, is played by unobserved LK types. If the type space is $K \backslash \{0\}$ and $\beta^{\kappa-1}(U) \notin S^{NE}$ then evolution from any interior initial condition converges to the asymptotically stable state where $x_{\kappa} = 1$.*

(b) *Suppose that the underlying game is a Travelers' Dilemma, played by unobserved LK types K. If $c \geq 2R + l - 1 + \kappa$ then evolution from any interior initial condition converges to an asymptotically stable state where only type κ exists.*

The condition $c \geq 2R + l - 1 + \kappa$ says that the strategy space is sufficiently large relative to the type space. For the case of $\kappa = 2$, $R = 3/2$, $P = 1/3$, and $c = 5$ part (b) is illustrated in figure 1.[13] The vertices (edges of the triangle) represent states where only one type exists. At the top vertex $x_0 = 1$, at the bottom left vertex $x_1 = 1$, and at the bottom right vertex $x_2 = 1$. The black dot denotes the asymptotically stable state.

Since only type κ survives in the WBRD-games, and since type κ earns the same as all types $k \geq 1$ in coordination games, the conclusion so far seems to be that evolution will wipe out all types except type κ. However, in cyclic games and in games with a unique interior ESS we get quite a different result. If the cyclic game is stable then it has a unique interior ESS, denoted σ^{ESS}. Recall that $\sigma(K, x)$ denotes the aggregate play in state x, and define the set of states where aggregate behavior corresponds to the unique interior ESS;

$$X^{ESS} = \left\{x \in \Delta(K) : \sigma(K, x) = \sigma^{ESS}\right\}.$$

Proposition 3. ***(a)*** *Suppose that an underlying cyclic game is played by unobserved LK types $K \backslash \{0\}$, with $\kappa \geq n$. All states with $\sum_{i \in \mathbb{N}} x_{k+i \cdot n} = 0$, for any $k \geq 1$,*

[13] All phase diagrams were created using the software Dynamo (Sandholm and Dokumaci 2007), with some additional editing in order to capture discontinuities, and asymptotically stable sets.

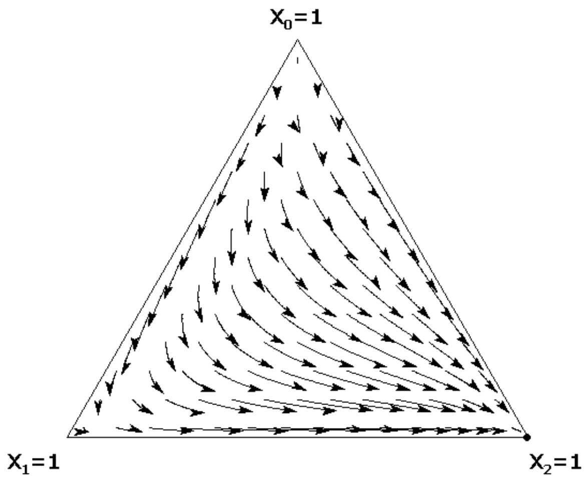

FIGURE 1. Travelers' Dilemma with unobserved LK types.

are unstable. If the underlying cyclic game is stable then evolution from any interior initial condition converges to X^{ESS}. The set X^{ESS} contains completely mixed states and if $\kappa = n$ then X^{ESS} is a singleton.

(b) *Suppose that the underlying game is a 2×2-game with a unique interior ESS, played by unobserved LK types K. No monomorphic states are stable and X^{ESS} is the unique asymptotically stable set, with the whole interior as its basin of attraction.*

The intuition for this result is that the payoffs in cyclic games are such that it is beneficial not to think, and behave, like everyone else. Alternatively one might point out that, as one approaches the ESS, the payoffs to different strategies are equalized, so that different types, playing different strategies, may earn the same. Figure 2, illustrates part (b) of the above proposition for the case of $\kappa = 2$ and $b = 2$. The thick line represent X^{ESS}.

The analysis so far has dealt with one game at a time. Now suppose that LK types are randomly matched to play games that are drawn from a set of games $\mathcal{G}$. If one disregard type 0 and lets the type space be $K\backslash\{0\}$ it is possible to prove results for the case when $\mathcal{G}$ consists of one game from each of the three general classes of a games; coordination games, cyclic games, and games satisfying WBRD. If type 0 is included in the analysis one can still prove results for the case when $\mathcal{G}$ consists of a 2×2 coordination game, a Hawk Dove game and a Travelers' Dilemma. The following results obtain:

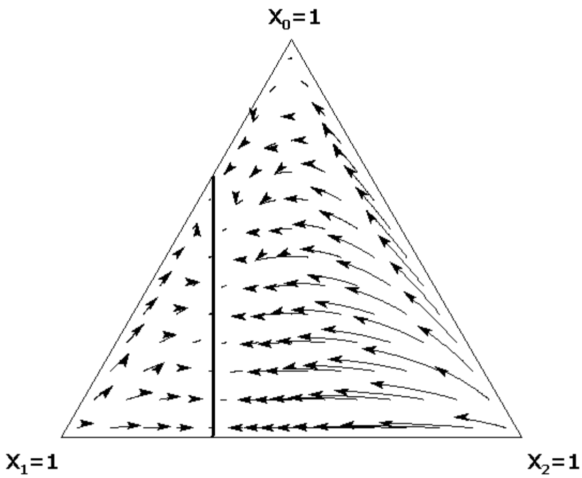

FIGURE 2. Hawk Dove with unobserved LK types.

PROPOSITION 4. ***(a)*** *Suppose that unobserved LK types* $K\backslash\{0\}$, *play games from a set* $\mathcal{G}$ *consisting of a coordination game, a cyclic game, and a game satisfying WBRD;*

$$\mathcal{G} = \{G^{Coord}, G^{Cyclic}, G^{WBRD}\}.$$

Suppose $\kappa = n$ *for the cyclic game and* $\beta^{\kappa-1}(U) \notin S^{NE}$ *for the WBRD-game. All monomorphic states are unstable if*

$$\frac{\mu^{Cyclic}}{\mu^{WBRD}} > \frac{w^{WBRD}(k,\kappa) - w^{WBRD}(1,\kappa)}{w^{Cyclic}(1,\kappa) - w^{Cyclic}(k,\kappa)},$$

for all $k \geq 2$. *And if the above inequalities is reversed for all* $k \geq 2$, *then only the state with* $x_\kappa = 1$ *is stable.*

(b) *Suppose that unobserved LK types* $K = \{0, 1, 2\}$ *play games from a set consisting of a* 2×2 *coordination game, a Hawk Dove game and a Travelers' Dilemma;*

$$\mathcal{G} = \{G^{2\times 2Coord}, G^{HD}, G^{TD}\},$$

with parameters $R = 3/2, P = 1/3,\ c \geq 4$. *If* $3\mu^{HD} > \mu^{TD}$ *then there is a unique asymptotically stable state, with the whole interior as basin of attraction, in which* $x_0 = 0$ *and*

$$x_1 = \frac{6b\mu^{HD} - 2\mu^{TD}}{6\mu^{HD}(b+1) + \mu^{TD}}.$$

If $3\mu^{HD} < \mu^{TD}$ *then only the state where* $x_2 = 1$ *is ESS.*

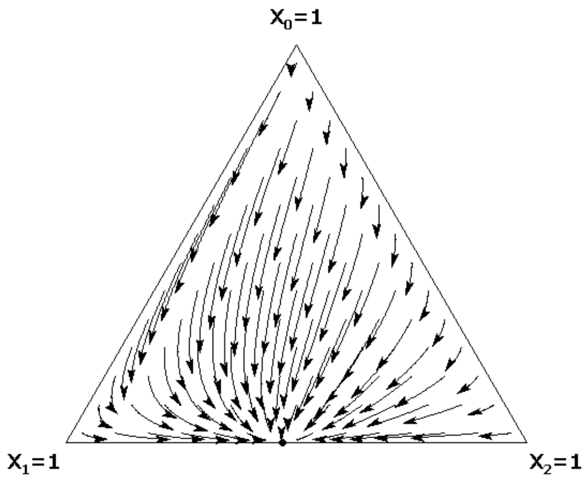

FIGURE 3. Evolution across games with unobserved LK types.

Since all types earn the same in coordination games, the dynamic is determined by the relations between the WBRD-game and the cyclic game in $\mathcal{G}$. In cyclic games type 1 has a payoff advantage over type κ when facing type κ. Conversely, in WBRD-games type κ has a payoff advantage over type 1 when facing type κ. Part (a) of the above proposition says that if the former advantage is sufficiently large, relative to the latter advantage, then all monomorphic states are unstable. The reason that the condition only involves type 1 and type κ is that the relations between the payoffs to the different types playing a cyclic game are very similar to the relations between the payoffs to the different types playing a WBRD-game. The only major difference lies in the payoff relations involving type 1 and type κ. Figure 3 illustrates part (b) of the above proposition for the case of $a = b = 1$, and $c = 5$. The black dot represents the unique asymptotically stable (and globally attracting) state where type 1 and 2 co-exist.

3.2. Partially Observed Types. For coordination games, the results from above are not altered when partial observation of types is introduced. It is still the case that all types $k \geq 1$ earn the same in all states, and all types $k \geq 1$ earn more than type 0. In the WBRD-games results may change drastically; it need no longer be the case that higher types earn more than lower types:

PROPOSITION 5. *Suppose that an underlying Travelers' Dilemma game, with $c \geq 6l + 1$, is played by partially observed LK types. If $P \in (p-1, p)$ for some $p \in \mathbb{N}$, then*

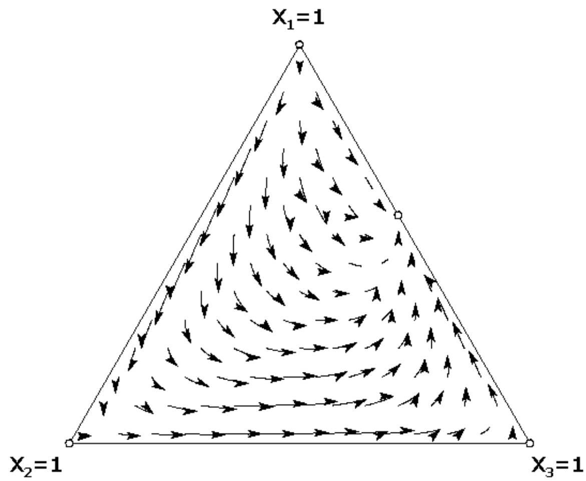

FIGURE 4. Travelers' Dilemma with partially observed LK types.

asymptotically $x_0 = 0$*, and every state where* $x_k = x_{k-1} = ... = x_{k-p} = 0$ *for some* $k > p$*, is unstable.*

The intuition for this result is that lower strategies are more destructive, so that when higher types meet each other they earn less, than what lower types earn when they meet higher types. Another way of putting this is to say that lower types have a committed advantage relative to higher types; a lower type is commited to a less destructive strategy, and may thereby induce a higher type to choose a less destructive strategy, something that might benefit both types. When there is a large fraction of the high type, this mechanism favors the growth of the low type. Figure 4 illustrates the above proposition for the case of $\kappa = 3$, $R = 3/2$, $P = 1/3$ and $c \geq 7$. Evolution from any interior initial state converges to the state $x = (4/7, 0, 3/7)$, which is not stable. The white dots represent unstable restpoints.

In cyclic games the results are also changed compared to the case of unobserved types:

PROPOSITION 6. *Suppose that an underlying cyclic game is played by partially observed LK types* $K \backslash \{0\}$*.*

(a) *If each strategy is the unique worst reply to itself, then all states where* $x_k = 0$ *for some* $k \geq 1$ *is unstable.*

(b) *If each strategy is the unique second best response to itself, then evolution from any interior initial condition converges to the state where* $x_\kappa = 1$*.*

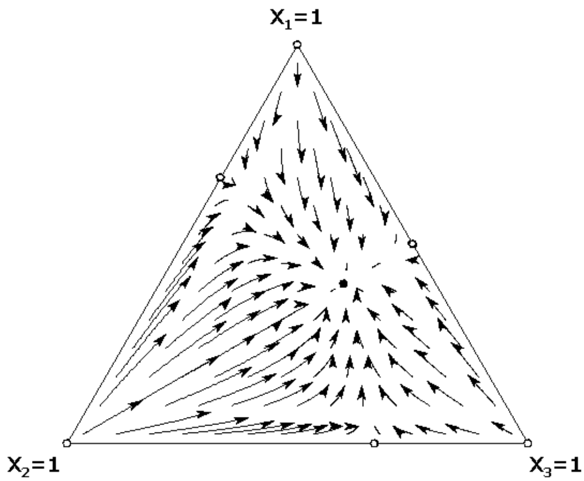

FIGURE 5. Hawk Dove with partially observed LK types.

(c) *Suppose that the underlying game is a 2×2-game with a unique interior ESS, played by partially observed LK types. If $\kappa \geq 3$ then evolution from any interior initial state converges to a unique interior state where $x_0 = 0$, and $x_i = bx_j$, for any odd number $i \leq \kappa$ and any even number $j \leq \kappa$.*

Part (a) says that for some cyclic games (those where each strategy is the unique worst reply to itself) introducing partial observability leads to an even stronger heterogeneity result than in the case of unobserved types. Before it was found that all states with $\sum_{i \in \mathbb{N}} x_{k+i \cdot n} = 0$ for some $k \geq 1$ are unstable. Now all states where $x_k = 0$ for some $k \geq 1$ are unstable. Like in the case of unobserved types there is a benefit of not thinking and behaving like everyone else. But now the lower types strengthen their position even further since opponents of higher types will understand what the lower type does, and thus the higher type rationally chooses to avoid a clash. In contrast to this, Part (b) says that for some other cyclic games (those where each strategy is the unique second best reply to itself) introducing partial observability leads to homogeneity. Part (c) strengthens the heterogeneity result of part (a) even further, for the case of 2×2-games. It establishes global convergence to a unique state where all types are present, except for type 0. This is illustrated in figure 5, for the case of $\kappa = 3$, and $b = 2$. Again, the black dot represents the asymptotically stable state and the white dots represent unstable restpoints.

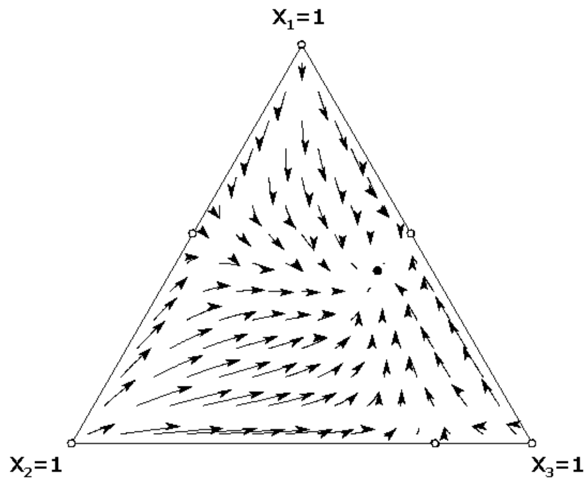

FIGURE 6. Evolution across games with partially observed LK types.

Suppose that partially observed LK types are randomly matched to play games that are drawn from a set of games $\mathcal{G}$ consisting of a 2×2 coordination game, a Hawk Dove game and a Travelers' Dilemma. The following results obtain:

PROPOSITION 7. *Suppose that partially observed LK types $K = \{0, 1, 2, 3\}$ play games from a set $\mathcal{G} = \{G^C, G^{HD}, G^{TD}\}$ with parameters $R = 3/2, P = 1/3, c \geq 7$. Let*

$$\hat{\mu} = \frac{\mu^{HD}}{\mu^{HD} + \mu^{TD}}.$$

If $\hat{\mu} < 1/4$ then evolution from any interior initial state converges to the state

$$x = \left(0, \frac{2(\hat{\mu}+2)}{5\hat{\mu}+7}, 0, \frac{3(\hat{\mu}+1)}{5\hat{\mu}+7}\right).$$

If $\hat{\mu} > 1/4$ then all states where some type $k \in \{1, 2, 3\}$ is extinct, are unstable.

Figure 6 below illustrates the dynamics for the case of $\mu^{HD} = \mu^{TD}$, and $a = b = 2$.

4. Extensions

4.1. Learning: Heterogeneous Fictitious Play.

4.1.1. *A Model of Heterogeneous Fictitious Play.* So far the focus has been on the evolution of theories of mind that are used to predict opponent's initial behavior. This section studies the evolution of theories of mind used in the process of learning. People may then use their information about past play to predict future play. Fictitious play postulates that all individuals believe that the future will be like the past, and best respond to the average of past play. I will modify this model and assume that there is a hierarchy of types $K = \{1, 2, ..., \kappa\}$, such that type 1 behaves in accordance with the fictitious play model, i.e. plays a best response to the average of past play, and type $k \geq 2$ plays a k times iterated best response to the average of past play. I will refer to this as the *heterogeneous fictitious play (HFP)* model. Note that there is *no type 0* in this model.

Suppose that during her lifetime each individual is randomly matched to play a symmetric two-player normal form game G (possibly drawn from a class of games) τ times with τ different individuals from the same population. The average payoff over these τ interactions serves as fitness payoff in the evolutionary process. In order to keep things tractable I assume that either all individuals have a common prior with full support, or they have (heterogeneous) priors which are formed according to the level-k model. Let $h^t \in \Delta(S)$ be the aggregate play in period t. According to fictitious play the belief γ^t evolves in the following way

$$\gamma^t = \frac{1}{t}\left(h^{t-1} + (t-1)\gamma^{t-1}\right).$$

Note that $\{h^t\}_{t=1}^{\tau}$ and $\{\gamma^t\}_{t=1}^{\tau}$ are fully determined by γ^1 and x. Type 1 plays strategy $\sigma(1, \gamma^t) = \beta(\gamma^t)$, and type k plays strategy

$$\sigma\left(k, \gamma^t\right) = \beta^k\left(\gamma^t\right).$$

If one allows for the possibility of partially observed types the behavior is as follows

$$\sigma\left(k, k', \gamma^t\right) = \begin{cases} \bar{\beta}^{k'+1}\left(\gamma^t\right) & \text{if } k' \leq k-1 \\ \bar{\beta}^{k}\left(\gamma^t\right) & \text{if } k' \geq k \end{cases}.$$

Regardless of whether types are unobserved or partially observed the expected payoff of type k, against type k', in period t, is

$$\Pi_k\left(k', \gamma^t\right) = \sum_{s\in S}\sum_{s'\in S} \sigma_s\left(k, k', \gamma^t\right)\sigma_{s'}\left(k', k, \gamma^t\right)\pi\left(s, s'\right).$$

Averaging over the τ periods, and recalling that $\{\gamma^t\}_{t=1}^{\tau}$ is fully determined by γ^1 and x, one gets

$$\Pi_k\left(k', x, \gamma^1\right) = \frac{1}{\tau}\sum_{t=1}^{\tau}\Pi_k\left(k', \gamma^t\right).$$

The expected payoff of type k, in state x is

$$\Pi_k\left(x, \gamma^1\right) = \sum_{k'\in K}\Pi_k\left(k', x, \gamma^1\right)x_{k'}.$$

This is the evolutionarily relevant payoff. The payoffs will generally not be continuous in the state. As a consequence, the vector field will generally not be Lipchitz continuous in the state, so we are unable to establish existence and uniqueness via the Picard-Lindelöf theorem. Therefore we will use the notion of a Filippov (1960) solution. (See Ito (1979) for a statement that is perhaps more accessible to economists.)

DEFINITION 8. *Consider the system $\dot{x} = \varphi(x)$ where φ is a real bounded measurable function, defined almost everywhere on a set $Q \subseteq \mathbb{R}^n$. Let*

$$C\{\varphi(x)\} = \cap_{\delta>0} \cap_{\mu(Z)=0} \overline{co}\{\varphi[\bar{B}_\delta(x) \setminus Z]\},$$

*where Z is an arbitrary set in $\mathbb{R}^n$, $\bar{B}_\delta(x)$ is the closed δ-ball around x, and $\overline{co}$ denotes the closed convex hull. A **Filippov solution** to the system $\dot{x} = \varphi(x)$ with initial condition x^0, is an absolutely continuous function $\xi(\cdot, x^0) : T \to Q$ with $\xi(0, x^0) = x^0$, such that*

$$\frac{\partial}{\partial t}\left(\xi\left(t, x^0\right)\right) \in C\{\varphi\left(\xi\left(t, x^0\right)\right)\},$$

holds almost everywhere.

Intuitively, the set $C\{\varphi(x)\}$ is constructed in the following way: From vectors associated with the ball $\bar{B}_\delta(x)$ we take away all those "strange" vectors that are only associated with some measure zero subset of the ball. Then we take the convex hull of the remaining non-strange vectors. Finally we take the limit as we shrink the ball. The useful thing about the Filippov solution is that it does not have to respect the direction of the vector field on a measure zero set. Filippov showed that a solution in the above sense always exists.

Our vector field $\varphi(x)$ satisfies the conditions of being, real, bounded, and measurable. Furthermore it is defined everywhere on $\Delta(K)$ except for the states with $x_0 = 0$. Hence our system has at least one Filippov solution. It turns out that this is all we need to prove the results we want.

4.1.2. *Unobserved Types.* In coordination games, all types behave in the same way, so any state is stable. For WBRD-games we have

PROPOSITION 8. *Suppose that an underlying game satisfying WBRD is played by unobserved HFP types. If $\beta(\gamma^1)$ is a singleton and if $\beta^{\kappa-1}(\gamma^1) \notin S^{NE}$ then evolution from any interior initial condition converges to the asymptotically stable state where only type κ exists.*

The assumption that $\beta^{\kappa-1}(\gamma^1) \notin S^{NE}$, guarantees that all types distinguish themselves behaviorally, at least in the first period. In cyclic games we have:

PROPOSITION 9. ***(a)*** *Suppose that an underlying cyclic game with n-strategies, is played by unobserved HFP types with $\kappa \geq n$, and suppose that $\beta(\gamma^1)$ is a singleton. All states with $\sum_{i \in \mathbb{N}} x_{k+i \cdot n} = 0$, for some k, are unstable.*

(b) *Suppose that the underlying game is a 2×2-game with a unique interior ESS, played by unobserved HFP types, and suppose that $\beta(\gamma^1)$ is a singleton. Let x_{odd} and x_{even} denote the fractions of odd and even types, respectively. No monomorphic states are stable and evolution from any interior initial state leads to the set where*

$$x_{odd}, x_{even} \in \left(\sigma_H^{ESS}, \sigma_D^{ESS}\right) = \left(1/(b+1), b/(b+1)\right).$$

The above result can be adapted to the following notorious example due to Shapley (1964);

$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

This game has a unique Nash equilibrium in which each of the three strategies are given equal weight. However, fictitious play does not converge to this equilibrium. Instead play goes round in a cycle whose time average does not correspond to the Nash equilibrium. Intuitively we might think that if real humans were engaged in this game and initially behaved in accordance with fictitious play, they would eventually be able to detect the cycles and best respond to it, and thereby break out of the cycle. The present model of heterogeneous fictitious play is able to do justice to these intuitions:

PROPOSITION 10. *Suppose that $\beta(\gamma^1)$ is a singleton. In the Shapley game evolution from any interior initial state converges to the state where*

$$\sum x_{1_{\bmod 3}} = \sum x_{2_{\bmod 3}} = \sum x_{3_{\bmod 3}} = 1/3,$$

and aggregate behavior corresponds to the unique interior Nash equilibrium $(1/3, 1/3, 1/3)$ *in all periods.*

4.1.3. *Partially Observed Types.* When partial observability is introduced into the *HFP* model the results are analogous to those obtained for partially observed *LK* types. In the Travelers' Dilemma we have:

PROPOSITION 11. *Suppose that an underlying Travelers' Dilemma game is played by partially observed HFP types. If $P \in (p-1, p)$ for some $p \in \mathbb{N}$, then asymptotically $x_0 = 0$, and every state where $x_k = x_{k-1} = ... = x_{k-1} = 0$ for some $k > p$, is unstable.*

In cyclic games:

PROPOSITION 12. *Suppose that an underlying cyclic game is played by partially observed HFP types.*

(a) *If each strategy is the unique worst reply to itself, then all states where* $x_k = 0$ *for some* k *is unstable.*

(b) *If each strategy is the unique second best response to itself, then evolution from any interior initial condition converges to the state where* $x_\kappa = 1$*.*

(c) *Suppose that the underlying game is a* 2×2*-game with a unique interior ESS, played by partially observed HFP types. If* $\kappa > 3$ *then evolution from any interior initial state converges to a unique interior state.*

4.2. A Nash Equilibrium Type. This section extends the level-*k (LK)* model by introducing a *Nash equilibrium (NE) type*, which is preprogrammed to play a Nash equilibrium strategy. Recall that in the Hawk Dove game there are three equilibria, two asymmetric and one symmetric, the latter one corresponding to an ESS. I will assume that the *NE* type plays the ESS strategy in these games. In coordination games there are two symmetric equilibria, both of which correspond to ESS. I this game I assume that the *NE* type chooses the equilibrium strategy that is the best reply against the uniform distribution over the strategy space. These assumptions are made in order to make the survival prospects of the *NE* type as good as possible. We will see that even under these favorable assumptions, the *NE* type does not dominate.

4.2.1. *A Nash Type Among Unobserved Level-k Types.* In the case of unobserved types I will assume that the *LK* types form beliefs as described previously, without taking the *NE* type into account. This assumption is made for three reasons: First, letting the *LK* types adjust their behavior to what the *NE* type would seem to bias the model in favor of the *LK* types. Second, since the *LK* model does not include the *NE* type, any assumption about how these types form beliefs about the *NE* type would be arbitrary. Third, from an evolutionary point of view, the question of whether a mutant *NE* type could take over a population of *LK* types should be well answered by examining the case when the *LK* types do not know about the *NE* type.

In coordination games the *NE* type plays $\beta(U)$, like type $k \geq 1$, so the fraction $x_{NE}/(1 - x_0 - x_{NE})$ will stay constant as evolution proceeds. In games with a unique interior ESS, the *NE* type plays the ESS so states with $x_{NE} > 0$ (including the state with $x_{NE} = 1$), as well as the state with $x_{NE} = 0$, belong to the set X^{ESS}. In the Travelers' Dilemma the results are more interesting, as described by the following proposition:

PROPOSITION 13. *Suppose that an underlying Travelers' Dilemma game is played by unobserved LK types and a NE type.*

(a) *If* $\kappa = c - l$ *then evolution from any interior initial condition converges to the asymptotically stable state where* $x_\kappa + x_{NE} = 1$*.*

(b) *If* $c - l - R < \kappa \leq c - l - 1$ *then evolution from any interior initial condition converges to the state where* $x_{NE} = 1$.

(c) *If* $\kappa < c - l - R$ *then evolution from any interior initial condition converges either to the state* $x_\kappa = 1$ *or the state* $x_{NE} = 1$*. Both of these states are asymptotically stable.*

If κ is large enough relative to c then case (a) or (b) applies. But if κ is not large enough relative to c, then case (c) applies, implying that evolution might lead to a state where the *NE* type is extinct. It is relevant to consider the case of a small κ, because it is plausible to assume that a *NE* mutant will not emerge until relatively sophisticated other types already are abundant. Intuitively we may say that it is counterproductive to think too much when your opponents do not think so much.

4.2.2. *A Nash Type Among Partially Observed Level-k Types.* In the case of partially observed types I will assume that the *LK* types do not take the existence of the *NE* type into account, so that the behavior of type k towards the *NE* type is the same as the behavior of type k towards higher types $k' > k$. I will assume that the *NE* type best responds to all other types, and when two *NE* individuals meet they play a Nash equilibrium – the same equilibrium as in the case of unobserved types. With these assumptions we have the following result in the Travelers' Dilemma:

Proposition 14. *Suppose that an underlying Travelers' Dilemma game with* $c \geq 6l+1$*, played by partially observed LK types, and a NE type. Suppose that* $P \in (p-1, p)$ *for some* $p \in \mathbb{N}$.

(a) *If* $\tilde{k} = \kappa$*, then asymptotically* $x_0 = 0$*, and every state where* $x_k = x_{k-1} = ... = x_{k-p} = 0$ *for some* $k > p$*, as well as the every state where* $x_\kappa + x_{NE} = x_{\kappa-1} = ... = x_{k-p} = 0$*, is unstable.*

(b) *If* $\tilde{k} \geq \kappa+1$*, then asymptotically* $x_0 = 0$*, and every state where* $x_k = x_{k-1} = ... = x_{k-p} = 0$ *for some* $k > p$*, as well as the every state where* $x_{NE} = x_\kappa = ... = x_{k-p} = 0$*, is unstable.*

That is, it is still the case that all monomorphic states are unstable. The *NE* type does not take over the population. Going over to the games with a unique interior ESS we find that all types co-exist.

Proposition 15. *Suppose that an underlying* 2×2*-game with a unique interior ESS, is played by partially observed LK types, and a NE type. If* $\kappa > 2$ *then evolution from any interior initial state converges to a unique interior state where* $x_0 = 0$*, and* $x_i = bx_j = -b/(1+b)\, x_{NE}$*, for any odd number* $i \leq \kappa$ *and any even number* $j \leq \kappa$.

4.3. The Cognitive Hierarchy Model . In the cognitive hierarchy (henceforth *CH*) model, individuals of type $k \geq 1$ believe that all other individuals belong to type

0 through $k-1$, and have a correct belief about the *relative population fractions* of lower types. One might ask how type k has arrived at a correct belief about the relative fractions of lower fractions. However, I leave this question aside and simply take the definition of types according to the CH model for granted. Type k believes that the population state is

$$\hat{x}^k = \frac{1}{\sum_{i=0}^{k-1} x_i} \left(x_0, x_1, ..., x_{k-1}, 0, ..., 0\right)'.$$

Note that in states where $x_0 = 0$, the beliefs and behavior of type 1 is not well-defined according to the model. Hence the beliefs of type $k \geq 1$ are also not well-defined. Similarly, if $x_1 = 0$ then the beliefs and behavior of types $k \geq 2$ are not well-defined. This is not an important limitation, since we are interested in evolution starting from an interior initial condition, and the system will not leave the interior in finite time. Let $\hat{\sigma}^k(K, x)$ be type k's belief about the population distribution of strategies, in state x, and let $\hat{\sigma}^k_s(K, x)$ be type k's belief about the share of the population that plays strategy s, in state x. Type k understands how lower types think and will behave. Hence

$$\hat{\sigma}^k_s(K, x) = \sum_{j=0}^{k-1} \sigma_s(j, x)\,\hat{x}_j.$$

Type 0 randomizes uniformly over the strategy space, i.e. plays $\beta(U)$. All types $k \geq 1$ best reply given their beliefs. Thus the behavior of type k at state x is given by

$$\sigma_k(x) = \bar{\beta}\left(\hat{\sigma}^k(K, x)\right).$$

The definitions of payoffs to types defined above are easily extended to this case where strategies depend on the state.

In the cognitive hierarchy model, the payoffs to different types will generally not be continuous in the state, so the vector field will generally not be Lipchitz continuous in the state. Therefore the notion of a Filippov solution will be used. The vector field $\varphi(x)$ satisfies the conditions of being, real, bounded, and measurable. Furthermore it is defined everywhere on $\Delta(K)$ except for the states with $x_0 = 0$. Hence the system has at least one Filippov solution.

The results concerning the *LK* are almost identical to those concerning the *CH* model, though the proofs are somewhat more complicated (see appendix). There is also a new result in proposition 18, to the effect that there is an asymptotically stable fraction of type 0 in some games. In coordination games it is trivial to see that the results are the same as in the *LK* model. In WBRD-games we have the follwing result:

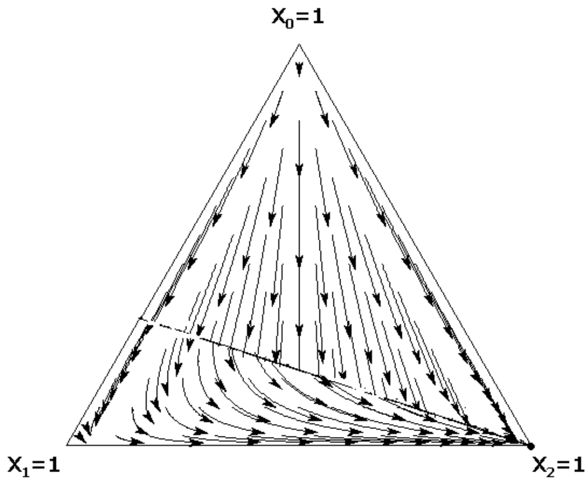

FIGURE 7. Travelers' Dilemma with (unobserved) CH types.

PROPOSITION 16. *Suppose that an underlying WBRD-game, is played by unobserved CH types. Evolution from any interior initial condition, following a Filippov solution, converges to the state where only type κ exists.*

For the Travelers' Dilemma, and the case of $\kappa = 3$, $R = 3/2$, and $P = 1/3$ this is illustrated in figure 7. Type 0 is extinct so the figure only depicts what happens when type 1, 2, and 3 are present. Type 1 plays the same strategy $\beta(U)$ in all states. Since $x_0 = 0$ in all states, type 2 plays the same strategy $\beta^2(U)$ in all states. Above the diagonal line type 3 plays the best response to what type 1 does, and below that line type 3 plays the best response to what type 2 does.

In the games with a unique interior ESS the only difference is that the set X^{ESS} looks somewhat different, but the general convergence result is the same.

PROPOSITION 17. *Suppose that an underlying 2×2-game with a unique interior ESS, is played by unobserved CH types. Evolution from any interior initial condition converges to X^{ESS}, which is the unique asymptotically stable set. No monomorphic states are stable.*

For the case of $K = \{0, 1, 2\}$, and $b = 1$, figure 8 illustrates the proposition. The thick (vertical and horizontal) lines represent X^{ESS}, the set of states where aggregate behavior corresponds to the ESS. In the area below (southwest of) the thin diagonal line type 2 plays H. Above this line type 2 plays D.

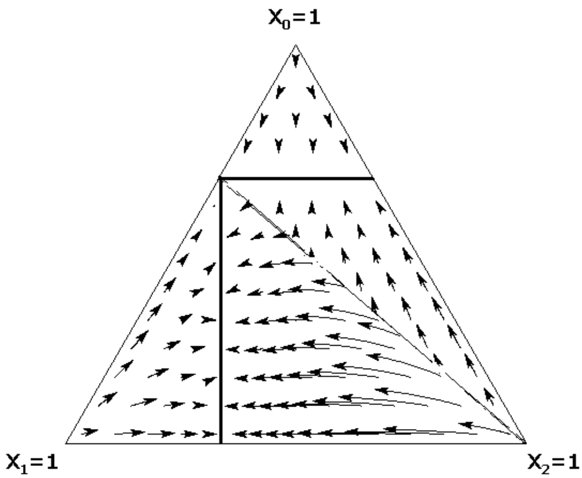

FIGURE 8. Hawk Dove with (unobserved) CH types.

In the games discussed so far, it is difficult for type 0 to survive. This need not be a problem, since it is usually found that type 0 is very rare. However, I will now analyze a game where type 0 has better chances of surviving. Let the game G^0 be defined by the following payoff matrix;

$$\begin{pmatrix} 3 & 2 & 0 \\ 8 & 0 & 0 \\ 3 & 3 & 1 \end{pmatrix}.$$

Name the strategies A , B, and C. Strategy A is strictly dominated by e.g. the mixed strategy that puts probability 1/4 on B and probability 3/4 on C. After deletion of strategy A, strategy C strictly dominates strategy B. Thus the remaining strategy profile (C, C) is the unique Nash equilibrium. We obtain the following result:

PROPOSITION 18. *Suppose that an underlying game G^0 is played by CH types. A state is Lyapunov stable if and only if it belongs to*

$$X^0 = \{x \in \Delta(K) : x_0 = 15/19\}.$$

No state is asymptotically stable and the set X^0 is the unique asymptotically stable set. The basin of attraction of X^0 includes states where x_0 is arbitrarily small.

For the case of $\kappa = 2$ figure 9 illustrates the dynamics. The thick (horizontal) line denotes the set X^0. Below the thin diagonal line type 2 plays C and above the thin line type 2 plays B. If type 0 and 1 is present in the population then aggregate behavior does not correspond to the Nash equilibrium, and if they are not present

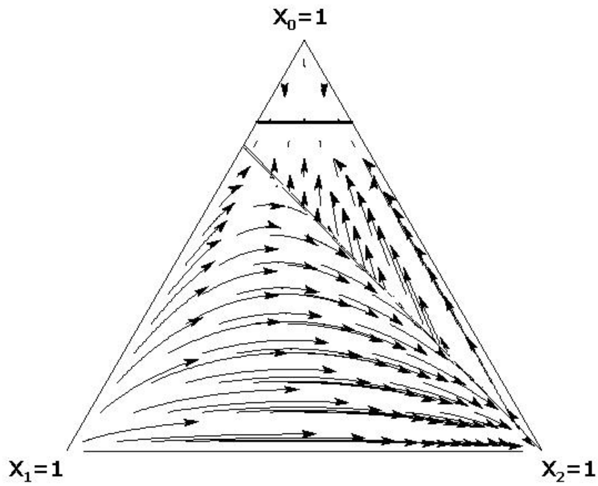

FIGURE 9. Evolution in game 0 with (unobserved) CH types.

then beliefs of type 2 are not well-defined. However, Nash equilibrium behavior is the limit of behavior as one move, from certain states below the diagonal, towards the state where everyone is of type 2 (where beliefs are not defined). Still this state is not attracting since evolution from some states arbitrarily close to the southwest corner (where everyone is of type 2) leads to the set X^0.

To understand what happens here, note that there is a region where everyone plays $\beta(U)$. Furthermore, note that the payoff that type 0 earns against an individual playing $\beta(U)$, is higher than the payoff that an individual playing $\beta(U)$ earns when meeting another individual doing the same thing. That is

$$\tilde{\pi}(U, \beta(U)) > \tilde{\pi}(\beta(U), \beta(U)).$$

So when the fraction of type 0 is small, and everyone except type 0 plays $\beta(U)$, then type 0 earns more than all other types. In general one can formulate the following sufficient condition for a set of states with $x_0 > 0$ to be asymptotically stable.

PROPOSITION 19. *Consider a finite symmetric two-player normal form game. Denote*

$$\tilde{\pi}(U, \beta(U)) - \tilde{\pi}(\beta(U), \beta(U)) = A,$$
$$\tilde{\pi}(\beta(U), U) - \tilde{\pi}(U, U) = B.$$

Suppose that the best reply to U is strict. There is some $\alpha \in (0,1)$ such that if $A/(A+B) > \alpha$, then the set of states where $x_0 = A/(A+B)$ is an asymptotically stable set.

4.4. Alternative Specification of Partial Observability . In this section I discuss an alternative way of modeling the case of partially observed types, in a way that captures the idea that lower types may understand that they are facing a higher type, although they do not know exactly how the higher type thinks. It is essentially a DK model (D for dominance) in the sense of Costa-Gomez and Crawford (2006), with the addition of partial observability.[14] This alternative model yields results which are almost identical to those obtained in the standard LK model with partially observed types.

Suppose two individuals A and B of different types k_A and k_B with $k_A < k_B$ are drawn to play a game. On the one hand, since B is of a higher type than A, I assume that B can observe A's type and understands how A thinks. Thus B plays a best response to what A does. On the other hand, since B is of a higher type than A, I assume that A does not understand how B thinks (even if A could observe B's type). Instead I only assume that A understands that since B is of a higher type, B will not play a strategy that A would never play. For evidence in favor of this general approach, see Palacios-Huerta and Volji (2009).

In order to make a sensible assumption about what it is that a player A would never do, one needs to think about why an individual forms beliefs in accordance with a certain type rather than another. I suggest that this is because it requires effort and reasoning power to entertain higher order beliefs – i.e. beliefs about beliefs, beliefs about beliefs about beliefs, and so on – and because this ability is heterogeneously distributed. Let a first order belief be a belief about some non-mental state; e.g. a belief that it snows. An organism is said to be first order intentional (Dennett 1987) if it is capable of forming beliefs about first order beliefs, e.g. able to form the belief that someone believes that it snows. An organism is second order intentional if it can form beliefs about beliefs about non-mental states, and so forth for higher order intentionality.[15] (Note that this means that being i^{th} order intentional is the same as being able to form $(i+1)^{th}$ order beliefs.)

[14] They find very weak support for the existence of this kind of cognitive types. However, in their set-up people have no specific information about the type of their opponent. So it is still an open question whether this kind of types can usefully describe the case of partially observed types.

[15] Kinderman et al. (1998) show that that normal humans find tasks involving more than fourth-order intentionality very hard (see also Apperly et al. 2007). The difficulties that higher order beliefs pose also matter in strategic settings: Kübler and Weizäcker (2004) estimate a quantal response model of beliefs in an information cascade experiment and find that the noise is increasing for higher order beliefs.

Type 1 only needs to form a first order belief about what type 0 will do. It is therefore reasonable to assume that type 1 is also first-order intentional. This means that type 1 can form a belief about its own belief, e.g. form the belief "I believe that type 0 will randomize uniformly". Type 2 needs to form a belief about what type 1 thinks that type 0 will do. Accordingly it is reasonable to assume that type 2 is second-order intentional, being able to form beliefs about what she believes that type 1 thinks that type 0 will do. Generally a type k individual needs to be able to entertain beliefs up to the k^{th} order about what other types believe. It is reasonable to assume that she is also able to reflect on the fact that she has these beliefs. Hence she is k^{th} order intentional.

Since a type 1 individual is able to form first order beliefs and best respond given these beliefs, type 1 will never play a first order dominated strategy. Since type 1 is first order intentional she is able to reflect on the fact that she will never play a first order dominated strategy. Consequently a type 1 individual should expect that an opponent of a higher type also does not play a first order dominated strategy. Similarly, a type k individual should expect that an opponent of type $k' > k$ does not play a k^{th} order dominated strategy. A type k individual does not know what a type $k' > k$ opponent chooses within the set of strategies that are not k^{th} order dominated. Therefore I will assume that type k follows the principle of insufficient reason and forms the belief that opponents of type $k' > k$ randomize uniformly over this set of strategies. The above line of reasoning also has implications for what assumption to make about what an individual of type k believes that other individuals of type k will do. Since she is unable to form more than k^{th} order beliefs, the best she can do is to form the same beliefs as she forms about higher type opponents.

Formally, define $D_0 = S$ and recursively define the set of strategies that are not i^{th} order dominated;

$$D_i = \{s \in S : s = \beta(t) \text{ for some } t \in D_{i-1}\}.$$

Against type k', type $k \geq 1$ plays

$$\sigma_k(k') = \begin{cases} \bar{\beta}(U) & \text{against } k' = 0 \\ \bar{\beta}^2(U(D_{k'})) & \text{against } k' \in \{1, 2, ...k-1\} \\ \bar{\beta}(U(D_k)) & \text{against } k' \in \{k, k+1, ...\kappa\} \end{cases}.$$

It should be noted that since $\bar{\beta}^2(U(D_{k'})) \in U(D_{k'})$, a lower type will never be surprised by what a higher type does. I will refer to this as the DK model.

For the Travelers' Dilemma only the restriction on c changes slightly compared to what we found before.

PROPOSITION 20. *Suppose that an underlying Travelers' Dilemma game with* $c \geq 4l + 7$, *is played by partially observed DK types. If* $P \in (p-1, p)$ *for some* $p \in \mathbb{N}$, *then asymptotically* $x_0 = 0$, *and every state where* $x_k = x_{k-1} = ... = x_{k-1} = 0$ *for some* $k > p$, *is unstable.*

For the games with a unique interior ESS, the only difference is that the globally attracting state is now symmetric:

PROPOSITION 21. *Suppose an underlying* 2×2*-game with a unique interior ESS is played by partially observed DK types. If* $\kappa > 2$ *then evolution from any interior initial state converges to the state where* $x_0 = 0$ *and* $x_k = 1/\kappa$ *for all* $k > 0$.

5. Discussion

There are only a few papers studying the evolution of cognitive types. A pioneering paper is Stahl (1993). In his model there is a set of types $n \in \{0, 1, 2, ...\}$ and all individuals are perfectly informed about the actual distribution of types in the population. Type 0 is divided into subtypes, each preprogrammed to different pure strategy. Type n believes that everyone else is of a lower type and is able to deduce what lower types will do. She chooses among strategies that are n^{th} order rationalizable conditional on the actual distribution of types. An individual of type n does not form any belief about what the opponent will choose among the set of n^{th} order rationalizable strategies. In order to choose among strategies in this set, each individual has a secondary strict preference ordering over strategies. Banerjee and Weibull (1995) study the interaction between individuals that are preprogrammed to different strategies and individuals that optimize given a correct belief about the strategy of the opponent (full information case) or the population distribution of strategies (incomplete information case). Another related paper is Stennek (2000) who studies the evolutionary advantage of ascribing different degrees of rationality to one's opponent. An individual of type $d \in \{0, 1, 2, ...\}$ believes that everyone else is of type $d - 1$ and chooses some c-iterations undominated action (because she is rational and also assumes that the opponents choose some $d - 1$-iterations undominated action). Like in Stahl's model the choice among the d-iterations undominated actions is made in accordance with some preference over the pure strategies rather than on the basis of a belief about which strategy (in the set) that the opponent will chose.

There are at least three important differences between the models in these papers and the present model: First, all of these papers assume that a fixed game is played recurrently and that some individuals are preprogrammed to pure strategies, like in the conventional evolutionary game theory literature. Since the payoff to different strategies varies from one game to another these models can not address the question of evolution of behavior in unfamiliar games (or games where there is little scope for learning). Secondly, these papers build on behavioral models that lack the kind of empirical support that the *LK* and *CH* models have (see Costa-Gomez and Crawford 2006 and Camerer 2003). Third, all of these models include many types whose behavior is not fully determined by best response given beliefs. Instead additional preference orderings are added in Stahl's and Stennek's studies, and several types are preprogrammed in the studies by Stahl and Bannerjee & Weibull. Such an approach potentially confounds the question of how theories of mind have evolved, and the question of how optimizing behavior has evolved. In contrast, in the *LK* and *CH* models, all types except level 0 best respond given their beliefs.

Sgroi and Zizzo (2009) study how a player endowed with a neural network adapts her initial play across games different 3×3-games, assuming that all other players play a Nash equilibrium strategy. Their simulation results are broadly consistent with a level-k model. As mentioned in the introduction there is almost no literature on evolution of different learning rules. Josephson (2008) uses simulations to compare fictitious play and reinforcement learning. He finds that evolution may end up putting roughly equal weights on these two modes of learning. Stahl's (2000) rule learning model assumes that each individual is endowed with propensities for different learning rules. One rule is to imitate past behavior, another one is to play a best response to past play. Further rules specify some iterated best response to past play. The propensities to follow these different rules are updated in relation to how well they perform in the game under study. This is undoubtedly an interesting model for explaining and predicting how people learn to play experimental games, but Stahl does not provide any general results on which propensities that may be evolutionarily stable in different classes of games.

Samuelson (2001a) studies the evolution of finite automata used to implement strategies in a class of three different games; an ultimatum game, an infinite horizon bargaining game, and a tournament. Each individual uses a single automaton to implement play in all three games. There is a cost that is increasing in the number of states of the automata. Thus there is an incentive to save on states and thereby behave in a way that is not tailored to each of the games – e.g. to behave in the same way in the ultimatum game as in the infinite horizon bargaining game. In a similar way, in the present paper I assume that the same theory of mind is used in all games.

This paper has identified mechanisms which can explain why different cognitive types may co-exist, most of which are relatively unsophisticated compared to the standards of rationalistic game theory. For unobserved types, in cyclic games, and in games with a unique interior ESS, we have seen that only mixed populations can be stable. The intuition behind this result is that in this kind of games the payoffs are such that it is advantageous not to behave like the opponent does. In the LK and CH models this translates into an advantage of not thinking like the opponent. The finding is similar to a result in Banerjee and Weibull (1995). They find that an optimizing type may earn less than a preprogrammed type in games with a unique interior ESS. However, their result only holds for the case of observable types, whereas we have found that lower types may have an advantage even when types are not observable.

Under partial observability in a Travelers' Dilemma it may be the case, that only polymorphic states are stable. The intuition for this result is that when their type can be observed, lower types may have a commitment advantage. This effect is reminiscent of results in the literature on preference evolution, where complete information about preferences can create similar commitment advantages (Dekel et al. 2006). In that context, assumptions about observability have sometimes been criticized for being ad hoc by Samuelson (2001b). However, in the context of the present models of theories of mind, the assumption about partial observability follows naturally from the fact that an individual's theory of mind is the individual's ability to understand how other people think.

From an evolutionary perspective, the potential advantage of a better theory of mind has to be weighted against the cost of increased reasoning capacity. Increased cognitive sophistication, in the form of higher order beliefs, is probably associated with non-negligible costs, see Holloway (1996), Dunbar (1998), and Apperly et al (2007). Such costs have been excluded from the formal analysis. However, the potential effect of cognitive costs should be kept in mind when interpreting the findings. Adding cognitive costs will increase the possibilities for lower types to survive.

According to the prominent "social brain", or "Machiavellian intelligence", hypothesis, the extraordinary cognitive abilities of humans, evolved as a result of the demands of social interactions, rather than the demands of the natural environment.[16] In a single person decision problem there is a fixed benefit of being smart, but in a strategic situation it may be important to be smarter than the opponent. Social interactions provided evolutionary incentives for the development of ever more advanced theories of minds among humans.

Robson (2003) models the Machiavellian intelligence hypothesis as the interaction between an uninformed player and an informed player. The informed does not want to

[16] Important references on the social brain, or Machiavellian, hypothesis include Jolly (1966), Humphrey (1976), Alexander (1990), Byren and Whiten (1998), Dunbar (1998), Dunbar (2003) and Flinn et al. (2005). Roth and Dicek (2005) discuss cross-species comparisons of cognitive abilities and conclude that humans have extraordinary cognitive abilities compared to other animals.

reveal her information, but the uninformed player wants her to do so. Both players use noisy bounded recall strategies. It turns out that for any equilibrium, each player would benefit from getting a longer recall. Thus there is a always pressure towards strategic sophistication in the form of greater recall. The results in this paper complement the social brain hypothesis, and Robson's results by suggesting mechanisms that may sustain heterogeneity with respect to theory of mind abilities (in a way that is consistent with experimental findings)

6. Conclusion

This paper makes three contributions. First, it performs an evolutionary analysis of the level-k and cognitive hierarchy models of initial play, as well as evolutionary analysis of a heterogeneous fictitious play model. Second, it extends these models to the case of partially observed types. Furthermore, the paper considers evolution of types across games in a novel way. It was found that an evolutionary process, based on payoffs earned in different games, both with and without partial observability, could lead to a polymorphic population where relatively unsophisticated types survive.

Appendix: Proofs

A1. Preliminaries.

Proof of Lemma 1. **(a)** Suppose that a game satisfies OSPP. Take any strategy $s < n$. By OSPP $s < \beta(s) \leq \beta^k(s)$ for all $k \geq 1$. By OSSP we also have that $\pi\left(\cdot, \beta^k(s)\right)$ is single peaked with its maximum at some $s' \geq \beta^k(s)$. Hence $\pi\left(\beta(s), \beta^k(s)\right) \geq \pi\left(s, \beta^k(s)\right)$.

Suppose that the game is strictly supermodular. For the setting of symmetric two player games the condition for a game to be strictly supermodular requires that the strategy space can be ordered $S = \{1, 2, ..., n\}$ such that the payoff function exhibits strictly increasing differences, in the sense that if $s > \tilde{s}$ and $s' > \tilde{s}'$ then

$$\pi(s, s') - \pi(\tilde{s}, s') > \pi(s, \tilde{s}') - \pi(\tilde{s}, \tilde{s}').$$

The set of equilibria has some smallest element $s^{NE}_{\min}$ and some largest element $s^{NE}_{\max}$. Vives (1990), theorem 5.1, shows that starting from $s < s^{NE}_{\min}$ ($s > s^{NE}_{\max}$) the Cournot best response dynamic converges monotonically upwards (downwards) to some point in S^{NE}. In a finite game this means that if $s < s^{NE}_{\min}$ then $s < \beta(s) \leq s^{NE}_{\max}$ and there exists some finite k such that $\beta^k(s) \in S^{NE}$. (Similarly, if $s > s^{NE}_{\max}$ then $s > \beta(s) \geq s^{NE}_{\max}$ and there exists some k such that $\beta^k(s) \in S^{NE}$.) Suppose $s < s^{NE}_{\min}$ so that $\beta(s) > s$ (the case with $s > s^{NE}_{\max}$ is exactly parallel). We have $\pi(\beta(s), s) - \pi(s, s) \geq 0$ and by supermodularity, it holds that

$$\pi(\beta(s), s') - \pi(s, s') \geq \pi(\beta(s), s) - \pi(s, s) \geq 0,$$

for all $s' > s$. In particular, since $\beta^k(s) \geq s$ for all $k \geq 1$ it holds that

$$\pi\left(\beta(s), \beta^k(s)\right) - \pi\left(s, \beta^k(s)\right) \geq 0,$$

for all $k \geq 1$. Finally note that if $k = 0$ then the above inequality is also satisfied.

(b) To see that strict supermodularity does not imply OSPP note that a supermodular game may have $\pi(s+2, s') > \pi(s+1, s') < \pi(s, s')$ for $s' < s$.

To see that OSPP does not imply supermodularity, note that the Travelers' Dilemma satisfies OSPP but is not supermodular, since if $s < \beta(s) < s'$, then we have $\pi(\beta(s), s) - \pi(s, s) > 0$, but $\pi(\beta(s), s') - \pi(s, s') = 0$. $\square$

In order to apply the LK and related models to the Travelers' Dilemma game we need the following lemma.

Lemma 2. *In the Travelers' Dilemma, the best reply to the uniform randomization over S is the strategy $s = c - l - 1$.*

Proof of Lemma 2. The expected payoff against the uniform randomization over a set $\{0, 1, ..., c\}$ is

$$\tilde{\pi}(s, U) = \frac{1}{c+1}\left(\sum_{s'=0}^{s-1}(s' - P) + s + \sum_{s'=s+1}^{c}(s+R)\right)$$
$$= \frac{1}{c+1}\left(s\left(\frac{(s-1)}{2} - P + 1\right) + (c-s)(s+R)\right).$$

A decrease from s to $s-1$ results in a change of payoff by

$$\tilde{\pi}(s-1, U) - \tilde{\pi}(s, U) = \frac{1}{c+1}\left((s-1)\left(\frac{(s-2)}{2} - P + 1\right) + (c-(s-1))((s-1)+R)\right)$$
$$- \frac{1}{c+1}\left(s\left(\frac{(s-1)}{2} - P + 1\right) + (c-s)(s+R)\right)$$
$$= \frac{1}{c+1}(P + R - c + s - 1).$$

This is weakly positive if $s \geq c+1-(P+R)$. So if one is currently using a sufficiently high strategy then it pays off to decrease one's strategy. Since we are assuming that $R + P = l + r$, for some $l \in \mathbb{N}$, and some $r \in (0, 1)$, we have

$$\beta(U) = \min\{s \in S : s > c - l - r + 1\}$$
$$= \min\{s \in S : s \geq c - l + 1\}$$
$$= c - l + 1.$$

□

It will be convenient to define a notion of *cognitive games*.

Definition 9. *Consider a symmetric two-player normal form game G and a set of cognitive types K. The corresponding* ***cognitive game*** *is a symmetric two-player game, where each player's strategy space is K, and the payoff to strategy k, against strategy k', is $w(k, k') = \Pi_k^G(k')$.*

Of course, for LK types, we have $\sigma(k) = \beta^k(U)$ for $k \geq 1$, and $\sigma(k) = U$ for $k = 0$. It may be instructive to draw an analogy to the notion of 'machine games' in the literature on games played by finite automata. One may view a cognitive game as depicting a situation where a fully rational principal has to delegate her strategy choice in game G to an agent who behaves in accordance with one of the types in K.

The definition of a cognitive game allows us to apply some results from standard evolutionary game theory, where evolution acts upon strategies, to the present setting where evolution acts upon the cognitive types. For instance, if a strategy in the cognitive game is ESS, then we know that the corresponding state $x \in \Delta(K)$ is asymptotically stable under the replicator dynamic.

A2. The Level-k Model with Unobserved Types.

Proof of Proposition 1. Trivial, therefore omitted. □

Before proving proposition 2 we establish the following lemma. We prove a more general result than we need here, but which we will need later.

Lemma 3. *Suppose that $\xi(t, x^0)$ is continuous everywhere, and differentiable in t almost everywhere, with*

$$\frac{\partial}{\partial t}\left(\xi\left(t, x^0\right)\right) = \left(\Pi\left(\xi\left(t, x^0\right)\right) - \bar{\Pi}\left(\xi\left(t, x^0\right)\right)\right).$$

If there exists a finite time t' such that for all $t \geq t'$ and all $i \in \{1, ..., k-1\}$ it holds that

$$\Pi_i\left(\xi\left(t, x^0\right)\right) - \Pi_{i-1}\left(\xi\left(t, x^0\right)\right) > \varepsilon_i,$$

for some $\varepsilon_i > 0$, then

$$\lim_{t\to\infty} \frac{\sum_{i=0}^{k-2} \xi_i\left(t, x^0\right)}{\xi_{k-1}\left(t, x^0\right)} = 0.$$

Proof of Lemma 3. Wherever $\xi(t, x^0)$ is differentiable we have

$$\begin{aligned}
&\frac{d}{dt}\left(\frac{\xi_{i-1}\left(t, x^0\right)}{\xi_i\left(t, x^0\right)}\right) \\
&= \frac{\dot{\xi}_{i-1}\left(t, x^0\right)\xi_i\left(t, x^0\right) - \xi_{i-1}\left(t, x^0\right)\dot{\xi}_i\left(t, x^0\right)}{\left(\xi_i\left(t, x^0\right)\right)^2} \\
&= \frac{\Pi_{i-1}\left(\xi\left(t, x^0\right)\right) - \bar{\Pi}\left(\xi\left(t, x^0\right)\right) - \left(\Pi_i\left(\xi\left(t, x^0\right)\right) - \bar{\Pi}\left(\xi\left(t, x^0\right)\right)\right)}{\left(\xi_i\left(t, x^0\right)\right)^2}\xi_i\left(t, x^0\right)\xi_{i-1}\left(t, x^0\right) \\
&= \left(\Pi_{i-1}\left(\xi\left(t, x^0\right)\right) - \Pi_i\left(\xi\left(t, x^0\right)\right)\right)\frac{\xi_{i-1}\left(t, x^0\right)}{\xi_i\left(t, x^0\right)} \\
&< -\varepsilon_i \frac{\xi_{i-1}\left(t, x^0\right)}{\xi_i\left(t, x^0\right)}.
\end{aligned}$$

Unless $\xi_{i-1}(t, x^0) / \xi_i(t, x^0) = 0$, the above expression is negative. Thus, since $\xi(t, x^0)$ is continuous everywhere we have

$$\lim_{t\to\infty} \frac{\xi_{i-1}\left(t, x^0\right)}{\xi_i\left(t, x^0\right)} = 0,$$

for all $i \in \{1, ..., k-1\}$. In a similar way have,

$$\begin{aligned}\frac{d}{dt}\left(\frac{\sum_{i=1}^{k-2}\xi_i\left(t,x^0\right)}{\xi_{k-1}\left(t,x^0\right)}\right) &= \sum_{i=1}^{k-2}\frac{d}{dt}\left(\frac{\xi_i\left(t,x^0\right)}{\xi_{k-1}\left(t,x^0\right)}\right)\\ &= \sum_{i=1}^{k-2}\left(\Pi_i\left(\xi\left(t,x^0\right)\right)-\Pi_{k-1}\left(\xi\left(t,x^0\right)\right)\right)\frac{\xi_i\left(t,x^0\right)}{\xi_{k-1}\left(t,x^0\right)}\\ &< -\sum_{i=1}^{k-2}\left(\varepsilon_i\frac{\xi_i\left(t,x^0\right)}{\xi_{k-1}\left(t,x^0\right)}\right),\end{aligned}$$

establishing the desired result. □

Proof of Proposition 2. The payoff matrix of the cognitive game is

$$\mathbf{A}=\begin{pmatrix} w(0,0) & w(0,1) & w(0,2) & . & w(0,\kappa)\\ w(1,0) & w(1,1) & w(1,2) & . & w(1,\kappa)\\ w(2,0) & w(2,1) & w(2,2) & . & w(2,\kappa)\\ . & . & . & . & .\\ w(\kappa,0) & w(\kappa,1) & w(\kappa,2) & . & w(\kappa,\kappa)\end{pmatrix}.$$

(a) (i) *Payoff relations*: By WBRD we have $w(k,i) \geq w(k-1,i)$ for all $i \geq k$. Since $\beta^{\kappa-1}(U) \notin S^{NE}$ we have $\beta^k(U) \neq \beta^{k-1}(U)$ for all $k \leq \kappa$, which implies $w(k,k-1) > w(k-1,k-1)$ for all $k \leq \kappa$.

(ii) *Convergence*: We proceed with a proof by induction. As inductive hypothesis suppose that there exists a finite time t^{k-1} such that for all $t \geq t^{k-1}$ and all $i \in \{2, ..., k-1\}$ it holds that

$$\Pi_i\left(\xi\left(t,x^0\right)\right)-\Pi_{i-1}\left(\xi\left(t,x^0\right)\right) > \varepsilon_i,$$

for some $\varepsilon_i > 0$. By lemma 3 we have

$$\lim_{t\to\infty}\frac{\sum_{i=1}^{k-2}\xi_i\left(t,x^0\right)}{\xi_{k-1}\left(t,x^0\right)}=0.$$

Note that

$$\begin{aligned}\Pi_k(x)-\Pi_{k-1}(x) &= \sum_{i=k}^{\kappa}\left(w(k,i)-w(k-1,i)\right)x_i\\ &\quad+\left(w(k,k-1)-w(k-1,k-1)\right)x_{k-1}\\ &\quad+\sum_{i=1}^{k-2}\left(w(k,i)-w(k-1,i)\right)x_i.\end{aligned}$$

The first term on the left hand side is non negative, the second term is strictly positive, and the third term may be positive or negative (in the Travelers' Dilemma it is negative). It follows that as $\sum_{i=1}^{k-2} x_i/x_{k-1} \to 0$, the above expression becomes strictly positive. Thus, there is some finite $t^k > t^{k-1}$ such that for all $t \geq t^k$ and all $i \in \{2, ..., k\}$ it holds that

$$\Pi_i\left(\xi\left(t, x^0\right)\right) - \Pi_{i-1}\left(\xi\left(t, x^0\right)\right) > \varepsilon_i > 0.$$

This completes the inductive step. The inductive base case is constituted by the observation that, when $x_0 = 0$, type 2 earns strictly more than type 1, in all states. We arrive at the result that there is some finite t^κ such that for all $t \geq t^\kappa$ it holds that

$$\Pi_i\left(\xi\left(t, x^0\right)\right) - \Pi_{i-1}\left(\xi\left(t, x^0\right)\right) > \varepsilon_i > 0,$$

for all $i \in \{2, ..., \kappa\}$. By lemma 3 this implies that evolution from any interior initial state converges to the state where $x_\kappa = 1$.

(iii) *Stability*: By WBRD, the second best response to $\beta^\kappa(U)$ is $\beta^{\kappa-1}(U)$ so all types $k < \kappa$, earn less than type κ against type κ. Therefore, the strategy κ in the cognitive game is ESS. Hence the state $x_\kappa = 1$ is asymptotically stable.

(b) Since the Travelers' Dilemma satisfies WBRD, it is sufficient to show that type 0 will become extinct. Since $\beta(U)$ is a pure strategy we have $\tilde{\pi}(\beta(U), U) = w(1, 0) > w(0, 0) = \tilde{\pi}(U, U)$. By lemma 2, $\beta(U) = c - l - 1$. It follows that, for $k \geq 1$;

$$\begin{aligned} w(1,k) - w(0,k) &= \tilde{\pi}(c-l-1, c-l-k) - \tilde{\pi}(U, c-l-k) \\ &= c - l - k - P \\ &\quad - \frac{1}{c+1}\left(\sum_{s=0}^{c-l-k-1}(s+R) + (c-l-k) + \sum_{s=c-l-k+1}^{c}(c-l-k-P)\right). \end{aligned}$$

Some algebra simplifies this expression to

$$\frac{1}{c+1}\left((c-l-k)\left(\frac{1}{2}(c-k-l+1) - R\right) + (c-k-l+1)P\right).$$

If $c \geq 2R + l - 1 + \kappa$ then $c \geq 2R + l - 1 + k$ for all k, which implies $c - k - l + 1 \geq 2R$ for all k. Furthermore this implies $c - k - l > 0$. Hence if $c \geq 2R + l - 1 + \kappa$, then $w(1, k) > w(0, k)$. Thus, type 0 is strictly dominated by type 1 in the cognitive game. □

Proof of Proposition 3. **(a)** Since $\sigma^{NE} \neq U$ the strategy $\beta(U)$ is unique. Thus if $\kappa = n$ each strategy will be played by exactly one type, so the payoffs of the cognitive game will be the same as in the underlying games, with the strategies renamed. If $\kappa \geq n$ then all strategies will be played by some type and some strategies will be played by more than one strategy. Thus the individuals of type k and type $k + n$ play the same strategy. It is easy to see that no monomorphic state is stable in a cyclic game. This implies that all states where there is some strategy that no type plays, are also unstable.

If the underlying game is stable, the cognitive game will also be stable. It is a standard result that the unique interior ESS in a stable game is globally attracting under the replicator dynamic, see e.g. Sandholm (2010).

(b) By the assumption about generic payoffs we have $\sigma^{ESS} \neq U$. Note the following property of 2×2-games with a unique interior ESS: If $\sigma_s(x) > \sigma_s^{ESS}$ then strategy $s \in \{H, D\}$ earns more than strategy $s' \neq s$. Let x_H, x_D, and x_U denote the fractions of the population that plays H, D, and U, respectively. In all interior state we have $x_H > 0$, $x_D > 0$, and $x_U > 0$. It is trivial to see that no monomorphic states are stable. Suppose the system is initially in an interior state where $x_H + x_U/2 > \sigma_H^{NE}$. Then $\dot{x}_H < 0$ and $\dot{x}_D > \dot{x}_U > \dot{x}_H$, so x_H decreases, x_D increases, and x_U may increase or decrease. This process continues until asymptotically $x_H + x_U = \sigma_H^{NE}$. Similar reasoning applies if the system is initially in an interior state where $x_D + x_U/2 > \sigma_D^{NE}$. Thus evolution from any interior initial state converges to some state where $x_s + x_U/2 = \sigma_s^{NE}$. $\square$

Proof of Proposition 4. **(a)** We study the cognitive game derived from the combination

$$G^{\mu,\mathcal{G}} = \mu^{Coord} \cdot G^{Coord} + \mu^{Cyclic} \cdot G^{Cyclic} + \mu^{WBRD} \cdot G^{WBRD}$$

All types earn the same in the coordination game so the dynamics is determined by the remaining two games.

In both the cognitive game derived from the cyclic game and the cognitive game derived from the WBRD-game it holds that strategy k is the unique best response to strategy $k-1$ for all $k \geq 2$. The best response functions of the two cognitive games differ only in that 1 is the unique best response to κ in the cognitive game derived from the cyclic game, whereas κ is the unique best response to κ in the cognitive game derived from the WBRD-game. This implies that type 1 is the unique best response to κ in the cognitive game derived from $G^{\mu,\mathcal{G}}$ if and only if

$$\begin{aligned} &\mu^{Cyclic} \cdot w^{Cyclic}(1,\kappa) + \mu^{WBRD} \cdot w^{WBRD}(1,\kappa) \\ &> \mu^{Cyclic} \cdot w^{Cyclic}(k,\kappa) + \mu^{WBRD} \cdot w^{WBRD}(k,\kappa), \end{aligned}$$

for all $k \geq 2$, or equivalently

$$\frac{\mu^{Cyclic}}{\mu^{WBRD}} > \frac{w^{WBRD}(k,\kappa) - w^{WBRD}(1,\kappa)}{w^{Cyclic}(1,\kappa) - w^{Cyclic}(k,\kappa)},$$

for all $k \geq 2$. If he above inequality holds, so that 1 is the unique best response to κ in the cognitive game derived from $G^{\mu,\mathcal{G}}$, then the this cognitive game is cyclic. Hence no monomorphic state is stable. If the above inequality is reversed for all $k \geq 2$ then κ is the unique best response to κ in the cognitive game derived from $G^{\mu,\mathcal{G}}$, so that the state $x_\kappa = 1$ is asymptotically stable.

(b) In the coordination game type 1 and 2 play $\beta(U)$. Thus the payoff matrix is

$$\begin{pmatrix} (a+1)/4 & a/2 & a/2 \\ a/2 & a & a \\ a/2 & a & a \end{pmatrix},$$

So type 0 is strictly dominated. In the Hawk Dove game type 1 plays $\beta(U) = D$, and type 2 plays $\beta(D) = H$. Thus the payoff matrix is

$$\begin{pmatrix} -(1+b)/4 & -1/2 & -b/2 \\ -1/2 & -1 & 0 \\ -b/2 & 0 & -b \end{pmatrix},$$

So type 0 earns the same as a mixed strategy in the cognitive game, which puts equal probability on type 1 and 2. In the Travelers' Dilemma type 1 plays $c-l-1$, and type 2 plays $c-l-2$. From lemma 2 we know that the payoff to strategy s against the uniform distribution is

$$\tilde{\pi}(s,U) = \frac{1}{c+1}\left(s\left(\frac{(s-1)}{2} - P\right) + s + (c-s)(s+R)\right).$$

Similarly, the payoff to uniform distribution against strategy s is

$$\tilde{\pi}(U,s) = \frac{1}{c+1}\left(s\left(\frac{(s-1)}{2} + R\right) + s + (c-s)(s-P)\right),$$

and the payoff when playing the uniform distribution against the uniform distribution is

$$\tilde{\pi}(U,U) = \frac{1}{c+1}\sum_{s=0}^{c}\tilde{\pi}(s,U) = \frac{1}{6}\frac{c}{c+1}(3R-3P+2c+1).$$

Using $P = 1/3$ and $R = 3/2$ yields the payoff matrix

$$\begin{pmatrix} \tilde{\pi}(U,U) & \tilde{\pi}(U,c-l-1) & \tilde{\pi}(U,c-l-2) \\ \tilde{\pi}(c-l-1,U) & c-2 & c-\frac{10}{3} \\ \tilde{\pi}(c-l-2,U) & c-\frac{3}{2} & c-3 \end{pmatrix}.$$

One can verify that type 0 is dominated for all $c \geq 4$.

The conclusion from these three games is that type 0 will be strictly dominated in the cognitive game based on a combination of games in $\mathcal{G}$, provided that $\mu^{HD} \neq 1$. Thus type 0 will be extinct so we disregard type 0 for the rest of the analysis. We can also disregard the coordination game since type 1 and 2 earn the same in that game. It follows that we can restrict attention to the following cognitive game between type

1 and 2, derived from the Hawk Dove game and the Travelers' Dilemma.

$$\begin{pmatrix} w(1,1) & w(1,2) \\ w(2,1) & w(2,2) \end{pmatrix} = \mu^{HD} \begin{pmatrix} -1 & 0 \\ 0 & -b \end{pmatrix} + \mu^{TD} \begin{pmatrix} -2 & -\frac{10}{3} \\ -\frac{3}{2} & -3 \end{pmatrix}$$
$$= \begin{pmatrix} -\mu^{HD} - 2\mu^{TD} & -\frac{10}{3}\mu^{TD} \\ -\frac{3}{2}\mu^{TD} & -b\mu^{HD} - 3\mu^{TD} \end{pmatrix}.$$

We have $w(1,1) < w(2,1)$ for all μ. We have $w(1,2) > w(2,2)$ if and only if $3b\mu^{HD} > \mu^{TD}$. In that case this game has a unique interor ESS where

$$x_1 = \frac{6b\mu^{HD} - 2\mu^{TD}}{6\mu^{HD}(b+1) + \mu^{TD}}.$$

If $3b\mu^{HD} < \mu^{TD}$ then only $x_2 = 1$ is ESS. □

A3. The Level-*k* Model with Partially Observed Types.

Proof of Proposition 5. Follows from lemma 4 and lemma 5. □

The following two lemmata are used in the proof of proposition 5:

Lemma 4. *In the Travelers' Dilemma, let $s(k,k')$ denote the pure strategy used by type $k \geq 1$ against type $k' \geq 1$. If type k' best responds to all types $k < k'$, and if*

$$s(1,k) > ... > s(k-1,k) > s(k-1,k) - 1 = s(k,k) = s(k,k+1) = ... = s(k,\kappa),$$

then every state where $x_k = x_{k-1} = ... = x_{k-p} = 0$ for some $k \geq 2$, is unstable.

Proof of Lemma 4. **(i)** *Payoff relations*: Suppose $k < k'$. Note that since $s(k,k') < s(k-1,k')$ we have

$$w(k,k') = s(k',k) - P = s(k,k') - 1 - P$$
$$< s(k-1,k') - 1 - P = s(k',k-1) - P = w(k-1,k').$$

Thus

$$w(1,k) > w(2,k) > ... > w(k-1,k).$$

Furthermore, it holds that

$$w(k-1,k) = s(k,k-1) - P = s(k-1,k) - 1 - P = s(k,k) - P = w(k,k) - P.$$

Moreover

$$w(k,k) = s(k,k) = s(k,k+1)$$
$$< s(k,k+1) - 1 + R = s(k+1,k) + R = w(k+1,k).$$

Finally, we have

$$w(k',k) = s(k',k) + R = s(k'+1,k) + R = w(k'+1,k),$$

which implies

$$w(k+1,k) = ... = w(\kappa,k).$$

In total we have found

$$\begin{aligned} w(1,k) - P > w(2,k) - P > ... > w(k-1,k) = \\ = w(k,k) - P < w(k,k) < w(k+1,k) = ... = w(\kappa,k). \end{aligned}$$

(ii) *Stability*: I prove the stability results only for the case of $P \in (0,1)$. Generalization is straightforward. We can disregard type 0 since all types $k \geq 1$ earn the same against type 0. Consider a state x where $x_i = x_{i-1} = 0$ for at least one $i > 0$. We have three different cases to consider: Either (I) there is at least one type $k < \kappa - 1$ such that $x_k = x_{k-1} = 0$ and $x_{k+1} > 0$, or (II) $x_{\kappa-1} = x_{\kappa-2} = 0$ and $x_\kappa > 0$, or (III) there is some $k \leq \kappa - 1$ such that $x_{k'} = 0$ for all $k' \geq k$, and $x_{k-1} > 0$.

Case I: Suppose that there is at least one type $k < \kappa - 1$ such that $x_k = x_{k-1} = 0$ and $x_{k+1} > 0$. The average payoff to types $k < \kappa - 1$ and $k + 1 < \kappa$ are, using $x_k = 0$,

$$\Pi_{k-1}(x) = \sum_{i=1}^{k-2} w(k-1,i)\, x_i + w(k-1,k+1)\, x_{k+1} + \sum_{i=k+2}^{\kappa} w(k-1,i)\, x_i,$$

and

$$\Pi_{k+1}(x) = \sum_{i=1}^{k-2} w(k+1,i)\, x_i + w(k+1,k+1)\, x_{k+1} + \sum_{i=k+2}^{\kappa} w(k+1,i)\, x_i,$$

so

$$\begin{aligned} \Pi_{k-1}(x) - \Pi_{k+1}(x) &= \sum_{i=1}^{k-2} \left(w(k-1,i) - w(k+1,i)\right) x_i \\ &+ \left(w(k-1,k+1) - w(k+1,k+1)\right) x_{k+1} \\ &+ \sum_{i=k+2}^{\kappa} \left(w(k-1,i) - w(k+1,i)\right) x_i. \end{aligned}$$

The first term on the right hand side is zero. The second term equals $(1-P)\, x_{k+1}$ so by the assumption that $P < 1$ this is positive. The third term is strictly positive, so $\Pi_{k-1}(x) > \Pi_{k+1}(x)$. Thus a mutant of type $k - 1 < \kappa - 2$ entering the population will earn more than type $k + 1 < \kappa$.

Case II: Suppose $x_{\kappa-1} = x_{\kappa-2} = 0$ and $x_\kappa > 0$. By a similar logic as in case I we have

$$\Pi_{\kappa-2}(x) - \Pi_\kappa(x) = \sum_{i=1}^{\kappa-3} \left(w(\kappa-2,i) - w(\kappa,i)\right) x_i + \left(w(\kappa-2,\kappa) - w(\kappa,\kappa)\right) x_\kappa,$$

where the first term on the right hand side is equal to zero, and the second term equals $(1-P)\,x_\kappa$. Thus a mutant of type $\kappa-2$ will earn more than type κ.

Case III: Suppose that there is some $k \le \kappa-1$ such that $x_{k'}=0$ for all $k' \ge k$, and $x_{k-1}>0$.The average payoff to type k is,

$$\Pi_k(x)=\sum_{i=1}^{k-1} w(k,i)\,x_i,$$

and the payoff to type $k-1$ is

$$\Pi_{k-1}(x)=\sum_{i=1}^{k-2} w(k-1,i)\,x_i + w(k-1,k-1)\,x_{k-1},$$

so

$$\begin{aligned}\Pi_k(x)-\Pi_{k-1}(x) &= \sum_{i=1}^{k-2}\left(w(k,i)-w(k-1,i)\right)x_i \\ &\quad + \left(w(k,k-1)-w(k-1,k-1)\right)x_{k-1}.\end{aligned}$$

The first term on the right hand side is zero and the second term is strictly positive, so $\Pi_k(x)>\Pi_{k-1}(x)$. Thus a mutant of type k entering the population will earn more than type $k-1$. □

LEMMA 5. *Consider the Travelers' Dilemma, played by partially observed LK types.*
(a) *For $k \ge 1$ the behaviors of the different types satisfy*

$$s(1,k)>...>s(k-1,k)>s(k-1,k)-1=s(k,k)=s(k,k+1)=...=s(k,\kappa).$$

(b) *If $c \ge 6l+1$ then type 1 earns strictly more than type 0 in all states.*

PROOF OF LEMMA 5. **(a)** By lemma 2 we have $\bar\beta(U)=c-l-1$. All types $k\ge 1$ play pure strategies as follows

$$s_k(k')=\begin{cases}\bar\beta(U)=c-l-1 & \text{against } k'=0\\ \bar\beta^{k'+1}(U)=c-l-k'-1 & \text{against } k'\in\{1,2,...k-1\}\\ \bar\beta^{k}(U)=c-l-k & \text{against } k'\in\{k,k+1,...\kappa\}\end{cases}.$$

It follows that $s(k,k)=s(k,k+1)=...=s(k,\kappa)$. Moreover, if $k'\ge k$ then $s(k,k')=c-l-k$ and $s(k-1,k')=c-l-(k-1)$, so

$$s(k-1,k')-s(k,k')=c-l-(k-1)-(c-l-k)=1.$$

Thus $s(1,k)>s(2,k)>...>s(k-1,k)$. Finally

$$s(k-1,k)-1=c-l-(k-1)-1=c-l-k=s(k,k).$$

(b) Since $\beta(U) = c - l - 1$ is a pure strategy we have $\tilde{\pi}(\beta(U), U) > \tilde{\pi}(U, U)$. Furthermore, for all $k' > 1$, we have $w(1,1) = c - l - 1 > c - l - 2 - P = w(1, k')$, so it is sufficient to show that $w(1, k') > w(0, k')$ for all $k' > 1$. In fact

$$\begin{aligned} w(1, k') - w(0, k') &= c - l - 2 - P - \tilde{\pi}(U, \beta(U)) \\ &= c - l - 2 - P \\ &\quad - \frac{1}{c+1}\left(\sum_{s=0}^{\beta(U)-1}(s+R) + \beta(U) + \sum_{s=\beta(U)+1}^{c}(\beta(U) - P)\right). \end{aligned}$$

Some algebra shows that this is equivalent to

$$\frac{1}{2(c+1)}\left(c^2 + c - 6cl - 2(2c - 2l - 1)r + 2R(c-l) + 5l^2 - l - 6\right).$$

Using $r \in (0,1)$, $c > l + 1/2$ and $R > 1$ (for the first inequality below) and $l > 1$ (for the second inequality below) we have

$$\begin{aligned} &2(c+1)(w(1,k') - w(0,k')) \\ &= \left(c^2 + c - 6cl - 2(2c - 2l - 1)r + 2R(c-l) + 5l^2 - l - 6\right) \\ &> \left(c^2 + c - 6cl - 2(2c - 2l - 1) + 2(c-l) + 5l^2 - l - 6\right) \\ &= c^2 - 6cl - c + 5l^2 + l - 4 \\ &> c^2 - 6cl - c + 2. \end{aligned}$$

This is positive if

$$c > 3l + \frac{1}{2}\sqrt{36l^2 + 12l - 7}.$$

It can easily be shown that

$$3l + \frac{1}{2}\sqrt{36l^2 + 12l - 7} + \frac{1}{2} < 6l + 1.$$

Thus if $c \geq 6l + 1$ then $w(1,1) > w(1,k') > w(0,k')$ for all $k' > 1$. □

Proof of Proposition 6. **(a)** Suppose that each strategy is the unique worst reply to itself. Then the payoff matrix of the underlying game is strategically equivalent to a game where the diagonal entries are zero, and all other entries are positive. For the cognitive game this implies that in each column (in the payoff matrix of the cognitive game), all entries below the diagonal are the same, and the diagonal entries are all zero. Furthermore, in each column of the cognitive game, the entries above the diagonal are strictly smaller than the entries below the diagonal (though all are positive). This implies that for all $k < \kappa$ it holds that

$$\begin{aligned}\Pi_k(x) - \Pi_{k+1}(x) &= \sum_{i=1}^{\kappa} \left(w(k,i) - w(k+1,i)\right) x_i \\ &= \left(w(k,k) - w(k+1,k)\right) x_k + \left(w(k,k+1) - w(k+1,k+1)\right) x_{k+1}.\end{aligned}$$

If $x_k = 0$ then, since the diagonal payoffs are zero, the above expression reduces to $w(k,k+1)\,x_{k+1}$, which is positive. Thus if $x_k = 0$ then a mutant of type k earns more than type $k + 1_{\mathrm{mod}\,n}$. (For $k = \kappa$ the above holds if one replaces $k+1$ with $\kappa + 1_{\mathrm{mod}\,n} = 1$.)

(b) Suppose that each strategy s is the unique second best response to itself (second only to the strategy $\beta(s) = s + 1_{\mathrm{mod}\,n}$). Then the payoff matrix of the underlying game is strategically equivalent to a game where the diagonal entries are zero, the entries corresponding to $\pi(s + 1_{\mathrm{mod}\,n}, s)$ are positive, and all other entries are negative. It follows that the payoff matrix of the cognitive game will have zeros on the diagonal, all entries below the diagonal positive, and all entries above the diagonal negative. Thus higher types strictly dominate lower types. Hence evolution from any interior initial condition converges to the state where $x_\kappa = 1$.

(c) *Behavior*: Let the first strategy be H and the second strategy be D. We have $\beta(U) = D$. Type $k \geq 1$ plays

$$\sigma_k(k') = \begin{cases} \bar{\beta}(U) = D & \text{against } k' = 0 \\ \bar{\beta}^{k'+1}(U) = H & \text{against odd } k' \in \{1,2,...k-1\} \\ \bar{\beta}^{k'+1}(U) = D & \text{against even } k' \in \{1,2,...k-1\} \\ \bar{\beta}^{k}(U) = D & \text{odd } k \text{ against } k' \in \{k,k+1,...\kappa\} \\ \bar{\beta}^{k}(U) = H & \text{even } k \text{ against } k' \in \{k,k+1,...\kappa\} \end{cases}.$$

Thus in each encounter between two different types $k > 0$ and $k' > 0$ ($k \neq k'$) the payoffs to both types are zero, and in encounters between two individuals of the same type, both earn -1 if they belong to an odd type and $-b$ if they belong to an even type. Finally we have $\tilde{\pi}(U,U) = -(1+b)/4$, and $\tilde{\pi}(\beta(U),U) = \tilde{\pi}(U,\beta(U)) = -1/2$. The payoff matrix for the cognitive game is

$$\begin{pmatrix} -(1+b)/4 & -1/2 & -1/2 & -1/2 & -1/2 & . \\ -1/2 & -1 & 0 & 0 & 0 & . \\ -1/2 & 0 & -b & 0 & 0 & . \\ -1/2 & 0 & 0 & -1 & 0 & . \\ -1/2 & 0 & 0 & 0 & -b & . \\ . & . & . & . & . & . \end{pmatrix}.$$

(ii) *Extinction of type 0*: Let $z = \{z_i\}_{i=1}^{\kappa}$ denote a mixed strategy in the cognitive game represented by the matrix above. Such a strategy z dominates type 0 in the cognitive game if and only if $z_i < 1/2$ for all odd i and $z_i < 1/2b$ for all even i. Thus, if κ is odd then there is a strategy z that dominates type 0 if and only if

$$\frac{1}{2}\frac{\kappa+1}{2}+\frac{1}{2b}\frac{\kappa-1}{2}>1,$$

or equivalently $\kappa > (3b+1)/(b+1)$. If κ is even then there is a strategy z that dominates type 0 if and only if

$$\frac{1}{2}\frac{\kappa}{2}+\frac{1}{2b}\frac{\kappa}{2}>1,$$

or equivalently $\kappa > 4b/(b+1)$. Since the former condition implies the latter condition, we conclude that if $\kappa > (3b+1)/(b+1)$, then type 0 is strictly dominated in the cognitive game. Thus the corresponding strategy in the cognitive game will asymptotically become extinct under the replicator dynamic. Finally note that the right hand side of this condition is increasing in b and then

$$\lim_{b\to\infty}\frac{3b+1}{b+1}=3.$$

Thus, for any finite b it holds that $(3b+1)/(b+1) < 3$.

(iii) *Convergence and stability*: After deletion of type 0 we have the payoff matrix

$$\mathbf{A}=\begin{pmatrix} -1 & 0 & 0 & 0 & . \\ 0 & -b & 0 & 0 & . \\ 0 & 0 & -1 & 0 & . \\ 0 & 0 & 0 & -b & . \\ . & . & . & . & . \end{pmatrix}.$$

It is clear that, regardless of whether κ is odd or even, we have $\Pi_1(x) = ... = \Pi_\kappa(x)$ if and only if $x_i = bx_j$, for any odd number $i \le \kappa$ and any even number $j \le \kappa$. This is the unique interior Nash equilibrium. In order to show that evolution from any interior initial state converges to a unique interior state it is sufficient to show that the game is stable. The tangent space is.

$$\mathbb{R}_0^\kappa = \left\{ v \in \mathbb{R}^\kappa : \sum_{i=1}^{\kappa} v_i = 0 \right\}.$$

A normal form game is stable if and only if the payoff matrix is negative definite with respect to the tangent space. Recall that the payoff matrix $\mathbf{A}$ is negative definite with respect to the tangent space if $v \cdot \mathbf{A}v < 0$, for all $v \in T(K\backslash\{0\})$, $v \neq \mathbf{0}$. One can transform the problem to one of checking negative definiteness with respect to the space $\mathbb{R}^{\kappa-1}$ rather than the tangent space. This is done with the following transformation matrix P (see Weissing 1991):

$$\mathbf{P}=\begin{pmatrix} 1 & 0 & . & 0 \\ 0 & 1 & . & 0 \\ . & . & . & . \\ 0 & 0 & . & 1 \\ -1 & -1 & . & -1 \end{pmatrix}.$$

Now check whether $(\mathbf{P}\cdot\mathbf{AP})$ is negative definite with respect to $\mathbb{R}^{\kappa-1}$. We have

$$\mathbf{P}\cdot\mathbf{AP}=\begin{pmatrix} -2 & -1 & -1 & -1 & . \\ -1 & -b-1 & -1 & -1 & . \\ -1 & -1 & -2 & -1 & . \\ -1 & -1 & -1 & -b-1 & . \\ . & . & . & . & . \end{pmatrix} = -\left(\mathbf{1}\cdot\mathbf{1}'\right)+\mathbf{I}\begin{pmatrix} -1 \\ -b \\ -1 \\ -b \\ \vdots \end{pmatrix},$$

where $\mathbf{1}$ denotes the column matrix with all entries equal to one and $\mathbf{I}$ denotes the identity matrix. Let $\mathbf{c}$ denote the vector with entries alternating between -1 and $-b$. Note that $v\cdot\left(-\left(\mathbf{1}\cdot\mathbf{1}'\right)\right)v<0$, for all $v\in\mathbb{R}^{\kappa-1}$, $v\neq\mathbf{0}$, so that $v\cdot(\mathbf{P}\cdot\mathbf{AP})\,v<0$ for all $v\in\mathbb{R}^{\kappa-1}$, $v\neq\mathbf{0}$, if and only if $\mathbf{Ic}$ is negative definite. Since $\mathbf{Ic}$ has two negative eigenvalues $-b$ and -1 it is indeed negative definite. This implies that $\mathbf{A}$ is negative definite with respect to the tangent space. □

Proof of Proposition 7. With $c\geq 7$, $R=3/2$, and $P=1/3$ ($\Rightarrow l=1$) and $\kappa\geq 3$ we have $c\geq 6+l$ so type 0 will become extinct in both the Travelers' Dilemma and the Hawk Dove game with partially observed types. In coordination games, type 0 is extinct for all a. Thus we can restrict attention to $K=\{1,2,3\}$. After excluding type 0, the payoffs in the cognitive game based on the Hawk Dove game are,

$$\mathbf{A}^{HD}=\begin{pmatrix} -1 & 0 & 0 \\ 0 & -b & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

For the Travelers' Dilemma we have

$$\begin{pmatrix} c-l-1 & c-l-2-P & c-l-2-P \\ c-l-2+R & c-l-2 & c-l-3-P \\ c-l-2+R & c-l-3+R & c-l-3 \end{pmatrix}.$$

Subtracting $c-l-3$ and using $R=3/2$, $P=1/3$ we get

$$\mathbf{A}^{TD}=\begin{pmatrix} 2 & \frac{2}{3} & \frac{2}{3} \\ \frac{5}{2} & 1 & -\frac{1}{3} \\ \frac{5}{2} & \frac{3}{2} & 0 \end{pmatrix}.$$

Putting together these two cognitive games we get

$$\hat{\mu}\mathbf{A}^{HD}+(1-\hat{\mu})\,\mathbf{A}^{TD}=\begin{pmatrix} 2-3\hat{\mu} & \frac{2}{3}(1-\hat{\mu}) & \frac{2}{3}(1-\hat{\mu}) \\ \frac{5}{2}(1-\hat{\mu}) & 1-b\hat{\mu}-\hat{\mu} & \frac{1}{3}(\hat{\mu}-1) \\ \frac{5}{2}(1-\hat{\mu}) & \frac{3}{2}(1-\hat{\mu}) & -\hat{\mu} \end{pmatrix}.$$

(i) Suppose $\hat{\mu} < 1/4$. This implies $w(3,3) > w(2,3)$. We also have $w(3,2) > w(2,2)$ for any $\hat{\mu}$ and b. Together this implies that type 2 earns less than type 3 in any state where $x_2, x_3 > 0$. Hence type 2 will be extinct, so we can restrict attention to the cognitive game between type 1 and 3, it is always the case that $w(3,1) > w(1,1)$ and $w(1,3) > w(3,3)$ so clearly there is always a unique interior ESS with $x_1 = (2\mu + 4)/(5\mu + 7)$, and $x_3 = 3(\mu + 1)/(5\mu + 7)$.

(ii) Suppose instead that $\hat{\mu} > 1/4$. This implies $w(3,3) < w(2,3)$ and $w(2,2) < w(1,2)$. We also have $w(2,3) < w(1,3)$, so in total the payoff relations are

$$\begin{aligned} w(3,1) &= w(2,1) > w(1,1) \\ w(3,2) &> w(2,2) < w(1,2) \\ w(3,3) &< w(2,3) < w(1,3). \end{aligned}$$

It is clear that no monomorphic state is stable. To verify that all states where $x_k = 0$ are unstable, first suppose that $x_1 = 0$, $x_2, x_3 > 0$. Type 1 earns more than type 2 in all such states so this is not stable. Second, suppose that $x_2 = 0$, $x_1, x_3 > 0$. Type 2 earns more than type 3 in all such states so this is not stable. Finally, suppose that $x_3 = 0$, $x_1, x_2 > 0$. Type 3 earns more than both type 1 and 2 in all such states so this is not stable. □

A4. Heterogeneous Fictitious Play. Behavior depends on the state and on initial beliefs, so let $w(k, k', x, \gamma^1) = \Pi_k(k', x, \gamma^1)$. The proofs from above regarding the *LK* model can be applied more or less directly to the *HFP* model, by using $w(k, k', x, \gamma^1)$ instead of $w(k, k')$ and showing that the payoff relations that obtain in the *LK* model, obtain for the *HFP* model, for almost every state x. The reason that we do not need to show that these relations obtain for measure zero sets of states is that we now study Filippov solutions.

Lemma 6. *For all games defined in this paper. If γ^1 has full support and if $\beta(\gamma^1)$ is a singleton, then the set of states x that, together with γ^1, induce histories such that some type is indifferent between two or more strategies in some period, has measure zero.*

Proof of Lemma 6. It is easy to verify that for all games defined in this paper it holds that if type 1 plays a pure strategy then higher types will play pure strategies. Thus it is sufficient to show that the set of states that induce histories in which type 1 is indifferent, has measure zero.

(i) Since γ^1 has full support, so has γ^t for all t. First we show that the set of beliefs with full support, at which type 1 indifferent between one or more strategies, has measure zero: The set of beliefs γ at which type 1 is indifferent between one or more strategies is

$$\Gamma^I = \{\gamma \in \Delta(S) : \gamma \text{ has full support and } \exists s, s^* \in S \text{ s.t. } \tilde{\pi}(s,\gamma) = \tilde{\pi}(s^*,\gamma)\}.$$

For all games defined in this paper it holds that for any pair of strategies s and s^* there is some s' such that $\tilde{\pi}(s,s') \neq \tilde{\pi}(s^*,s')$. This implies that the dimension of Γ^I is lower than the dimension of $\Delta(S)$. Thus Γ^I is a hyperplane with measure zero.

(ii) Now we show that the set of states that induce histories where type 1 is indifferent, has measure zero. Recall $\gamma^t = (h_{t-1} + (t-1)\gamma^{t-1})/t$. Since Γ^I is a hyperplane it follows that if $\gamma^{t-1} \notin \Gamma^I$ then there is a measure zero set of states that induce aggregate behavior h_{t-1} such that $\gamma^t \in \Gamma^I$. Since we have assumed that $\beta(\gamma^1)$ is a singleton, an inductive argument establishes that, for each period, the set of states that induce indifference, given γ^1, has measure zero. Since the number of periods is finite it follows that the set of states that induce indifference in some period, given γ^1, has measure zero. □

Proof of Proposition 8. By lemma 6 we can assume that all types play pure strategies in all states, without loss of generality. Let $s(k,\gamma^t)$ denote the pure strategy that is chosen by type k given the belief γ^t. Higher types play weakly lower strategies, i.e. for all γ^t it holds that $k > k'$ implies $s(k,\gamma^t) \leq s(k',\gamma^t)$. Since $\beta^{\kappa-1}(\gamma^1) \notin S^{NE}$ we have $s(k,\gamma^t) < s(k',\gamma^t)$ at least for $t = 1$. This implies that the following payoff relations hold, analogous to what was obtained for the LK model above:

If $k < k''$ and $k' < k''$ then

$$\pi\left(\sigma\left(k,\gamma^t\right),\sigma\left(k'',\gamma^t\right)\right) = \pi\left(\sigma\left(k',\gamma^t\right),\sigma\left(k'',\gamma^t\right)\right),$$

for all t, so summing over the τ periods we have

$$w\left(k,k'',x,\gamma^1\right) = w\left(k',k'',x,\gamma^1\right).$$

Moreover, we have

$$\pi\left(\sigma\left(k,\gamma^t\right),\sigma\left(k-1,\gamma^t\right)\right) \geq \pi\left(\sigma\left(k-1,\gamma^t\right),\sigma\left(k-1,\gamma^t\right)\right),$$

for all t, with strict inequality as long as type $k-1$ plays a strategy above zero, so summing over the τ periods we have

$$w\left(k,k-1,x,\gamma^1\right) > w\left(k-1,k-1,x,\gamma^1\right).$$

Furthermore, for all $i \leq k-2$ we have

$$\pi\left(\sigma\left(k,\gamma^t\right),\sigma\left(i,\gamma^t\right)\right) \leq \pi\left(\sigma\left(k-1,\gamma^t\right),\sigma\left(i,\gamma^t\right)\right),$$

for all t, with strict inequality as long as type $k-1$ plays a strategy above zero, so summing over the τ periods we have

$$w\left(k, i, x, \gamma^1\right) < w\left(k-1, i, x, \gamma^1\right).$$

The rest of the proof is identical to the proof of proposition 2. □

Proof of Proposition 9. Since $\beta\left(\gamma^1\right)$ is a singleton, by lemma 6 we can assume that all types play pure strategies in all states.

(a) The proof is very similar to the proof of proposition 3, and therefore omitted.

(b) We can divide the population into odd and even types. In each period t, all odd types play $\bar{\beta}\left(\gamma^t\right)$, and all even types play $\bar{\beta}^2\left(\gamma^t\right)$. Suppose that the fraction of odd types, is very small, so that the strategy they play will always be underweighted relative to the ESS, i.e. $x_{odd} < \sigma_H^{ESS} < \sigma_D^{ESS}$. Hence they will earn more than the even types in all periods, and the fraction of odd types will therefore grow. Eventually a state is reached in which $\sigma_H^{ESS} < x_{odd} < \sigma_D^{ESS}$. In such a state x_{odd} may grow or decline depending on the relative frequency at which the different strategies are played by the different types. Similarly, if we start in a state with $x_{even} < \sigma_H^{ESS}$, we will eventually reach a state where $\sigma_H^{ESS} < x_{even} < \sigma_D^{ESS}$. Using $\sigma_H^{ESS} = 1/\left(1+b\right)$ and $\sigma_D^{ESS} = b/\left(1+b\right)$ we conclude that evolution from any interior initial state leads to the set where $x_{odd}, x_{even} \in \left(1/\left(b+1\right), b/\left(b+1\right)\right)$. (If one of the fractions belong to this interval then so does the other.) □

Proof of Proposition 10. Since $\beta\left(\gamma^1\right)$ is a singleton, by lemma 6 we can assume that all types play pure strategies in all states. In period t type 1 plays $\beta\left(\gamma^t\right)$ and type k plays $\beta\left(\gamma^t\right) + \left(k+1\right)_{\text{mod}\,3}$. Thus in each period the payoffs to types $1_{\text{mod}\,3}$, $2_{\text{mod}\,3}$ and $3_{\text{mod}\,3}$ are given by the matrix

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix},$$

where the entry in the i^{th} row and the j^{th} column represents the payoff to type $i_{\text{mod}\,3}$ against type $j_{\text{mod}\,3}$. Note that this matrix is the transpose of the payoff matrix of the underlying game. It is a standard result that the replicator dynamics converges from any interior initial state to the unique state $(1/3, 1/3, 1/3)$. □

Proof of Proposition 11. For a given t let $s\left(k, k', \gamma^t\right)$ denote the pure strategy used by type $k \geq 1$ against type $k' \geq 1$. In each period t type k plays

$$\max\left\{\bar{\beta}^k\left(\gamma^{t-1}\right), 0\right\} = \max\left\{\bar{\beta}\left(\gamma^{t-1}\right) + 1 - k, 0\right\}$$

against type $k' \geq k$, and plays a best response to lower types. Thus for each t we have

$$\begin{aligned} s\left(1,k,\gamma^t\right) &\geq ... \geq s\left(k-1,k,\gamma^t\right) \geq s\left(k-1,k,\gamma^t\right)-1 \\ &= s\left(k,k,\gamma^t\right) = s\left(k,k+1,\gamma^t\right) = ... = s\left(k,\kappa,\gamma^t\right), \end{aligned}$$

with all the inequalities being strict, at least in the first round ($t = 1$). It follows that the payoff relations that obtained *strictly* in the LK model (see proof of lemma 4) now obtain *weakly* in all periods. Moreover, they obtain strictly at least in the first period. Hence, when we take the average over all periods, the same strict payoff relations obtain as in the LK model. The rest of the proof follows the proof of lemma 4. □

Proof of Proposition 12. The proof of part (a) and (b) is very similar to the proof of proposition 6a and 6b, and are therefore omitted.

(c) In each period t any odd type k plays $\bar{\beta}(\gamma^t)$ against type $k' \geq k$, and any even type k plays $\bar{\beta}^2(\gamma^t)$ against type $k' \geq k$. They best respond to lower types. It follows that in each period t, type k earns a payoff of zero against all opponents except when facing an individual of the same type k, in which case she earns -1 or $-b$. Averaging over periods is straightforward gives us a payoff matrix (for each state x) of the cognitive game, which is a diagonal matrix with a negative diagonal. We can use the proof of proposition 6 to establishing negative definiteness with respect to the tangent space. □

A5. A Nash Equilibrium Type.

Proof of Proposition 13. The proof is the same as that of proposition 2 up to the result that there is some finite t^κ such that for all $t \geq t^\kappa$ it holds that $\Pi_1(\xi(t,x^0)) < ... < \Pi_\kappa(\xi(t,x^0))$. This implies that evolution from any interior initial state leads to some state where $\sum_{i=0}^{\kappa-1} x_i = 0$. Thus, for the rest of the proof we only need to consider the cognitive game with types κ and *NE*. There are three different cases to consider:

(a) If $\kappa = \tilde{k}$ (or equivalently $\kappa = c - l$) then $w(\kappa, NE) = w(NE,NE) = w(NE,\kappa) = w(\kappa,\kappa)$. Thus evolution from any $x^0 \in int(\Delta(K))$ leads to some state where $x_\kappa + x_{NE} = 1$.

Suppose that $\kappa = \tilde{k} - j$, for some $j \geq 1$. In this case $w(\kappa, NE) = -P$, $w(NE,NE) = 0$, $w(NE,\kappa) = R$, and $w(\kappa,\kappa) = j \geq 1$. That is, the payoff matrix in the cognitive game between type κ and type NE is

$$\begin{pmatrix} j & -P \\ R & 0 \end{pmatrix}$$

(b) If $j < R$ (or equivalently $\tilde{k} - R < \kappa \leq \tilde{k} - 1$) then evolution from any $x^0 \in int(\Delta(K))$ leads to the state where $x_{NE} = 1$.

(c) If $j > R$ (or equivalently $\kappa < \tilde{k} - R$) then evolution from any $x^0 \in int\left(\Delta\left(K\right)\right)$ leads either to the state $x_\kappa = 1$ or the state $x_{NE} = 1$. Both of these states are asymptotically stable. □

Proof of Proposition 14. The proof of proposition 2, via lemma 4, can be adapted. □

Proof of Proposition 15. Adding a *NE* type to the payoff matrix from the cognitive game in the proof of proposition 6 we have;

$$\begin{pmatrix} -\left(b+1\right)/4 & -1/2 & -1/2 & -1/2 & . & -1/2 \\ -1/2 & -1 & 0 & 0 & . & 0 \\ -1/2 & 0 & -b & 0 & . & 0 \\ -1/2 & 0 & 0 & -1 & . & 0 \\ . & . & . & . & . & . \\ -1/2 & 0 & 0 & 0 & . & -b/\left(1+b\right) \end{pmatrix}.$$

Let $z = \{z_i\}_{i=1}^{\kappa} \cup \{z_{NE}\}$ denote weights put on the types $k \geq 1$ and type NE in the cognitive game represented by the matrix above. Such a strategy z dominates type 0 in the cognitive game if and only if $z_i < 1/2$ for all odd i, and $z_i < 1/2b$ for all even i, and if $z_{NE} < \left(1+b\right)/2b$. Thus, if κ is odd then there is a strategy z that dominates type 0 if and only if

$$\frac{1}{2}\frac{\kappa+1}{2} + \frac{1}{2b}\frac{\kappa-1}{2} + \frac{1+b}{2b} > 1,$$

or equivalently $\kappa > 2b/\left(1+b\right)$. If κ is even then there is a strategy z that dominates type 0 if and only if

$$\frac{1}{2}\frac{\kappa}{2} + \frac{1}{2b}\frac{\kappa}{2} + \frac{1+b}{2b} > 1,$$

or equivalently $\kappa > 2\left(b-1\right)/\left(b+1\right)$. Since the former condition implies the latter condition, we conclude that if $\kappa > 2b/\left(1+b\right)$, then type 0 is strictly dominated in the cognitive game. Finally note that the right hand side of this condition is increasing in b and

$$\lim_{b\to\infty}\frac{2b}{1+b} = 2.$$

After deletion of type 0 it is clear that we have $\Pi_1\left(x\right) = ... = \Pi_\kappa\left(x\right)$ if and only if $x_1 = bx_2 = x_3 = bx_4 = ... = -b/\left(1+b\right)x_{NE}$. The rest of the proof is the same as for proposition 6. □

A6. The Cognitive Hierarchy Model. Behavior depends on the state, so let $w\left(k,k',x\right) = \Pi_k\left(k',x\right)$. The proofs from above regarding the *CH* model can be modified to apply to the *CH* model, by using $w\left(k,k',x\right)$ instead of $w\left(k,k'\right)$ and showing that the payoff relations that obtain in the *LK* model, obtain for the *CH* model, for almost every state x. Like for the *HFP* model, the reason that we do not need to show that these relations obtain for measure zero sets is that we now study Filippov solutions.

Proof of Proposition 16. As noted in the text, we know that there exists at least one Filippov solution through any initial state. Since we do not know if the solution is unique, we have to prove that any solution through an interior initial state converges to the state where $x_\kappa = 1$.

(a) Consider any interior initial state x^0 and any Filippov solution $\xi(\cdot, x^0)$ through x^0. To show that $\xi(\cdot, x^0)$ convergence to $x_\kappa = 1$, we proceed with a proof by induction: As inductive hypothesis suppose that there exists a finite time t^{k-1} such that for all $t \geq t^{k-1}$ and all $i \in \{2, ..., k-1\}$ it holds that $\sigma_i(\xi(t, x^0)) = \beta^i(U)$ and that

$$\Pi_i\left(\xi\left(t, x^0\right)\right) - \Pi_{i-1}\left(\xi\left(t, x^0\right)\right) > \varepsilon_i,$$

for some $\varepsilon_i > 0$. (In addition to this it always holds that $\sigma_1(\xi(t, x^0)) = \beta(U)$.) By lemma 3 we have

$$\lim_{t\to\infty} \frac{\sum_{i=1}^{k-2} \xi_i(t, x^0)}{\xi_{k-1}(t, x^0)} = 0.$$

This implies that asymptotically type k believes that everyone is of type $k-1$; thus

$$\lim_{t\to\infty} \sigma_k\left(\xi\left(t, x^0\right)\right) = \beta^k(U).$$

Moreover, note that when $\sigma_i(\xi(t, x^0)) = \beta^i(U)$ for all $i \in \{2, ..., k\}$, then

$$\begin{aligned}
\Pi_k(x) - \Pi_{k-1}(x) = &\sum_{i=k}^{\kappa} (w(k, i, x) - w(k-1, i, x))\, x_i \\
&+ (w(k, k-1, x) - w(k-1, k-1, x))\, x_{k-1} \\
&+ \sum_{i=1}^{k-2} (w(k, i, x) - w(k-1, i, x))\, x_i.
\end{aligned}$$

Like in the proof of proposition 2, the first term on the left hand side is nonnegative, the second term is strictly positive, and the third term is positive or negative. Thus, there is some finite $t^k > t^{k-1}$ such that for all $t \geq t^k$ and all $i \in \{2, ..., k\}$ it holds that $\sigma_i(\xi(t, x^0)) = \beta^i(U)$ and that

$$\Pi_i\left(\xi\left(t, x^0\right)\right) - \Pi_{i-1}\left(\xi\left(t, x^0\right)\right) > \varepsilon_i,$$

for some $\varepsilon_i > 0$. This completes the inductive step.

The inductive base case is constituted by the observation that when $x_0 = 0$ we have $\sigma_2(\xi(t, x^0)) = \beta^2(U)$ for all t, and type 2 earns strictly more than type 1, in all states.

We arrive at the result that there is some finite t^{κ} such that for all $t \geq t^{\kappa}$ and all $i \in \{2, ..., \kappa\}$ it holds that $\sigma_i(\xi(t, x^0)) = \beta^i(U)$ and that

$$\Pi_i\left(\xi\left(t, x^0\right)\right) - \Pi_{i-1}\left(\xi\left(t, x^0\right)\right) > \varepsilon_i > 0.$$

By lemma 3 this implies that $\xi(\cdot, x^0)$ converges to the state where $x_\kappa = 1$.

(b) The proof of proposition 2 can be adapted to show that type 0 will become extinct. To see this, note that for each state x, the difference $w(1, k, x) - w(0, k, x)$ in the *CH* model, will equal the difference $w(1, k') - w(0, k')$ for some k' in the *LK* model. Thus type 0 earns strictly less than type 1 in all states, so that $C\{\varphi(x)\}$ only contains vectors such that $\dot{x}_0 < \dot{x}_1$. It follows that $x_0 \to 0$. □

Proof of Proposition 17. **(i)** *Rest points*: First we show that a state $x \in \Delta(K)$ is a rest point of the replicator dynamics if and only if it is monomorphic or belongs to X^{ESS}: To see that all points in X^{ESS} are rest points, note that all strategies in the support of a Nash equilibrium earn the same payoff against the Nash equilibrium strategy. Since the unique ESS is interior, all strategies earn the same against σ^{ESS}. Hence if $x \in X^{ESS}$ then all types earn the same, so x is a rest point. Furthermore, it is trivial that monomorphic states are rest points.

To see that all rest points are either monomorphic or in X^{ESS}, consider a polymorphic state x where $\sigma(K, x) \neq \sigma^{ESS}$. Behavior cannot correspond to any of the two asymmetric Nash equilibria, so $\sigma(K, x) \neq \sigma^{ESS}$ implies $\sigma(K, x) \neq \sigma^{NE}$, which means that one strategy earns more than the other. If the lowest type randomizes uniformly then the second lowest type plays a pure strategy – since we have assumed $\sigma^{ESS} \neq U$. These two types earn different payoffs so the state is unstable.

(ii) *Unstable states*: Now we show that no monomorphic states are stable: To see that a state with $x_k = 1$, $k \in \{1, ..., \kappa - 1\}$, is unstable, note that there is some $\varepsilon > 0$ such that in any state with $x_k \in (1 - \varepsilon, 1)$, type k plays $\bar{\beta}\left(\hat{\sigma}^k(K, x)\right)$ and all types $k' > k$ play a different strategy $\bar{\beta}^2\left(\hat{\sigma}^k(K, x)\right)$. If $x_k \in (1 - \varepsilon, 1)$ we have $\Pi_{k'} \geq -b(1 - x_k)$ for all $k' > k$, and $\Pi_k \leq -x_k$. We have $-b(1 - x_k) > -x_k$ if and only if $x_k > b/(1 + b)$. Thus there is some $\delta \leq \varepsilon$ such that if $x_k \in (1 - \delta, 1)$, and $x_{k'} > 0$, then $\Pi_{k'} > \Pi_k$ for all $k' > k$. This means that if the system starts in such a point it moves away from the monomorphic state where $x_k = 1$.

A similar argument shows that the state with $x_0 = 1$ is unstable: To see that the state with $x_\kappa = 1$ is unstable let $\delta = \min_{i \in \{H,D\}} \sigma_i^{ESS}$. Since σ^{ESS} is interior we have $\delta > 0$. Assume, without loss of generality, that $\beta(U) = D$. If $x_\kappa = 1 - \varepsilon$, and $x_0 = \varepsilon$, then type κ plays $\beta(U) = D$. In this state $\sigma_H(x) = 1 - \varepsilon/2$ and $\sigma_D(x) = \varepsilon/2$. If $\varepsilon < \delta$ then $\sigma_H(x) = \varepsilon/2 < \varepsilon < \delta = \min_{i \in \{H,D\}} \sigma_i^{ESS} \leq \sigma_H^{ESS}$. Thus H earns a higher payoff than D against $\sigma(K,x)$. It follows that type 0 earns more than type κ in all states where $x_0 < \delta$.

(iii) *Stable states***:** Now we show that X^{ESS} is the unique asymptotically stable set, with the whole interior as its basin of attraction: If the system starts in $x^0 \in X^{ESS}$ then the system remains in this set, so assume $x^0 \notin X^{ESS}$. Let X^I denote the set of states where one or more types $k \geq 1$ are indifferent between the strategies;

$$X^I = \left\{x \in \Delta(K) : \exists k \text{ s.t. } \hat{\sigma}^k(K,x) = \sigma^{ESS}\right\}.$$

The set X^I is closed, since a type is indifferent between strategies only when they yield the exact same expected payoff. Let $int(\Delta(K))$ denote the interior of $\Delta(K)$.

(iii.i) Suppose that $x^0 \in int(\Delta(K))$, but $x^0 \notin X^I$. Since X^I is closed, there is a neighborhood B of x^0 such that in every state $x \in B$ all types use the same strategy as in x^0. Since $x^0 \notin X^{ESS}$ one strategy i is overweighted relative to its weight in the ESS, i.e. $\sigma_i(x^0) > \sigma_i^{ESS}$. This implies that strategy i earns less than strategy $j \neq i$. Thus the fractions of the types that play strategy i decrease as the system moves away from x^0. A type k that initially plays strategy i does so because it mistakenly believes that strategy i is underweighted relative to its weight in the ESS, i.e. $\hat{\sigma}_i^k(x^0) < \sigma_i^{ESS}$. As the fractions of all types playing strategy i decrease, it continues to hold that $\hat{\sigma}_i^k(x) < \sigma_i^{ESS}$, so no type that plays i switches to j. There may be some types that start out by playing j which eventually come to believe that strategy i is underrepresented relative to the ESS (since the fraction that plays i decreases). Thus either (1) the fraction of types playing pure strategy i decreases until a state in X^{ESS} is reached, or (2) the fraction of types playing pure strategy i goes to zero, and the fraction of type 0 (which put probability 1/2 on strategy i) decreases until a state in X^{ESS} is reached.

(iii.ii) Suppose that $x^0 \in int(\Delta(K)) \cap X^I$. Since $x^0 \notin X^{ESS}$ and $x^0 \in int(\Delta(K))$ (i) implies that evolution will lead away from x^0, and that the fractions of all types will change (including type 0). This will change the beliefs of all types $k \geq 2$, so that they are no longer indifferent between strategies. Thus the system moves away from X^I, and the rest follows from (iii.i). □

Proof of Proposition 18. The fact that X^0 is an asymptotically stable set follows from proposition 19. The rest of the (tedious) proof is available from the author upon request. □

Proof of Proposition 19. Note that $B \geq 0$ so $A/(A+B) > 0$ implies $A > 0$. Since the best reply to U is strict, there exists an $\alpha \in (0,1)$ such that if $x_0 \geq \alpha$ then

$$\beta(U) = \arg\max_{\sigma \in \Delta(S)} \sigma \cdot A\left(\hat{\sigma}^k(K,x)\right),$$

for all k. Thus if $x_0 \geq \alpha$ then type $k \geq 1$ play $\beta(U)$ so that all types $k \geq 1$ earn the same payoff. For all $k > 0$ we have

$$\begin{aligned}\Pi_0 - \Pi_k &= x_0 \tilde{\pi}(U, U) + (1 - x_0) \tilde{\pi}(U, \beta(U)) \\ &- (x_0 \tilde{\pi}(\beta(U), U) + (1 - x_0) \tilde{\pi}(\beta(U), \beta(U))) \\ &= x_0 (\tilde{\pi}(U, U) - \tilde{\pi}(\beta(U), U)) \\ &+ (1 - x_0)(\tilde{\pi}(U, \beta(U)) - \tilde{\pi}(\beta(U), \beta(U))) \\ &= -Bx_0 + (1 - x_0) A.\end{aligned}$$

This is positive if and only if $A/(A + B) > x_0$ (implying $A > 0$). Hence if $A/(A + B) > x_0 > \alpha$ then $\Pi_0 > \Pi_k$ for all $k > 0$, and if $x_0 > A/(A + B) > \alpha$ then $\Pi_0 < \Pi_k$ for all $k > 0$. □

A7. Alternative Specification of Partial Observability.

PROOF OF PROPOSITION 20. Follows from lemma 4 and lemma 7. □

LEMMA 7. *In the Travelers' Dilemma, with partially observed DK types, all types $k \geq 1$ use pure strategies.*

(a) *For $k \geq 1$ the behaviors of the different types satisfy*

$$s(1, k) > ... > s(k - 1, k) > s(k - 1, k) - 1 = s(k, k) = s(k, k + 1) = ... = s(k, \kappa).$$

(b) *If $c \geq 4l + 7$ then type 1 earns strictly more than type 0 in all states.*

PROOF OF LEMMA 7. **(a)** We have $D_k = \{0, 1, ..., c - k\}$, for all $k \geq 0$, so by lemma 2 it holds that $\bar{\beta}(U(D_k)) = c - k - l - 1$. Thus all types $k \geq 1$ play pure strategies as follows

$$s_k(k') = \begin{cases} c - l - 1 & \text{against } k' = 0 \\ c - k' - l - 2 & \text{against } k' \in \{1, 2, ...k - 1\} \\ c - k - l - 1 & \text{against } k' \in \{k, k + 1, ...\kappa\} \end{cases}.$$

It follows that $s(k, k) = s(k, k + 1) = ... = s(k, \kappa)$. Moreover, if $k' \geq k$ then $s(k, k') = c - k - l - 1$ and $s(k - 1, k') = c - (k - 1) - l - 1$ so

$$s(k - 1, k') - s(k, k') = c - (k - 1) - l - 1 - (c - k - l - 1) = 1.$$

Thus $s(1, k) > s(2, k) > ... > s(k - 1, k) = s(k, k)$.

(b) An argument very similar to that in the proof of lemma 5 establishes that if $c > 4l + 7$ then $w(1, 1) > w(1, k') > w(0, k')$ for all $k' > 1$. □

PROOF OF PROPOSITION 21. As before, let the first strategy be H and the second strategy be D. Suppose without loss of generality that $a > b$, so that $\beta(U) = D$. Type

$k \geq 1$ plays

$$\sigma_k(k') = \left\{ \begin{array}{cc} \bar{\beta}(U) = D & \text{against } k' = 0 \\ \bar{\beta}^2(U) = H & \text{against } k' \in \{1, 2, ...k-1\} \\ \bar{\beta}(U) = D & \text{against } k' \in \{k, k+1, ...\kappa\} \end{array} \right. .$$

Thus in each encounter between two different types $k > 0$ and $k' > 0$ $(k \neq k')$ the payoff is zero, and in encounters between two individuals of the same type, both earn -1. Moreover, we have $\tilde{\pi}(U, U) = -(1+b)/4$, and $\tilde{\pi}(\beta(U), U) = \tilde{\pi}(U, \beta(U)) = -1/2$, so the payoff matrix for the cognitive game is

$$\begin{pmatrix} -(1+b)/4 & -1/2 & -1/2 & . \\ -1/2 & -1 & 0 & . \\ -1/2 & 0 & -1 & . \\ . & . & . & . \end{pmatrix}.$$

A mix between the types above 0 strictly dominates type 0 if and only if $-1/\kappa > -1/2$, or equivalently $\kappa > 2$. After deletion of type 0 from the above matrix, what remains is $-\mathbf{I}$. The rest of the proof follows the same logic as the proof of proposition 6. □

References

Alexander, R. D. (1990), 'How did Humans Evolve? Re‡ections on the Uniquely Unique Species', *University of Michigan Museum of Zoology Special Publication* No 1.

Apperly, I. A., Back, E., Samson, D. and France, L. (2007), 'The Cost of Thinking about False Beliefs: Evidence from Adult's Performance on a Non-Inferential Theory of Mind Task', *Cognition* 106, 1093–1108.

Banerjee, A. and Weibull, J. W. (1995), Evolutionary Selection and Rational Behavior, in A. Kirman and M. Salmon, eds, *Learning and Rationality in Economics*, Blackwell, Oxford, UK, chapter 12, pp. 343–363.

Basu, K. (1994), 'The Travellers'Dilemma: Paradoxes of Rationality in Game Theory', *American Economic Review* (Papers and Proceedings) 84(2), 391–395.

Brown, A., Camerer, C. and Lovallo, D. (2008), 'To Review or Not To Review? Limited Strategic Thinking at the Movie Box Office', Working paper, Texas AM University.

Byrne, R. W. and Whiten, A. (1998), *Machiavellian intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans*, Oxford University Press, Oxford.

Camerer, C. F. (2003), *Behavioral Game Theory*, Princeton University Press, Princeton.

Camerer, C. F., Ho, T.-H. and Chong, J.-K. (2004), 'A Cognitive Hierarchy Model of Games', *Quarterly Journal of Economics* 199, pp. 861–898.

Coricelli, G. and Nagel, R. (2009), 'Neural Correlates of Depth of Strategic Reasoning in Medial Prefrontal Cortex', *Proceedings of the National Academy of Sciences* USA 106(23), 9163–9168.

Cosmides, L. and Tooby, J. (1992), 'Cognitive Adaptations for Social Exchange', in J. Barkow, L. Cosmides and J. Tooby, eds, *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, Oxford University Press, New York.

Costa-Gomes, M. A. and Crawford, V. P. (2006), 'Cognition and Behavior in Two-Person Guessing Games: an Experimental Study', *American Economic Review* 96, 1737–1768.

Crawford, V. (2003), 'Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions', *American Economic Review* 93, 133–149.

Crawford, V. and Iriberri, N. (2007), 'Level-k Auctions: Can Boundedly Rational Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?', *Econometrica* 75, 1721–1770.

Dekel, E., Ely, J. C. and Yilankaya, O. (2007), 'Evolution of Preferences', *Review of Economic Studies* 74, 685–704.

Dennett, D. C. (1987), *The Intentional Stance*, MIT Press, Cambridge, Massachusetts.

Dunbar, R. I.M. (1998), 'The Social Brain Hypothesis', *Evolutionary Anthropology* 6, 178–190.

Ellingsen, T. and Östling, R. (2009), When Does Communication Improve Coordination? forthcoming in the *American Economic Review*.

Filippov, A. F. (1960), 'Differential Equations with Discontinuous Right-Hand Side', *Matematicheskii Sbornik* 51, 199–231. English translation, 1964, *American Mathematical Society Translations Series* 2, 42, pp. 199-231.

Fudenberg, D. and Levine, D. K. (1998), *The Theory of Learning in Games*, MIT Press, Cambridge, MA.

Goeree, J. K. and Holt, C. A. (2001), 'Ten Little Treasures of Game Theory and Ten Intuitive Contradictions', *American Economic Review* 91(5), 1402–1422.

Goeree, J. K. and Holt, C. A. (2004), 'A model of noisy introspection', *Games and Economic Behavior* 46(2), 365–382.

Haruvy, E. and Stahl, D. O. (2009), 'Learning Transference Between Dissimilar Symmetric Normal-Form Games', Working paper, University of Texas at Austin.

Holloway, R. (1996), 'Evolution of the Human Brain', in A. Lock and C. R. Peters, eds, *Handbook of Human Symbolic Evolution*, Clarendon, Oxford, pp. 74–125.

Humphrey, N. K. (1976), 'The social function of intellect', in P. P. G. Bateson and R. A. Hinde, eds, *Growing Points in Ethology*, Cambridge University Press, Cambridge, pp. 303–317.

Ito, T. (1979), ‘A Filippov solution of a system of differential equations with discontinuous right-hand sides’, *Economics Letters* 4(4), 349–354.

Josephson, J. (2008), ‘A numerical analysis of the evolutionary stability of learning rules’, *Journal of Economic Dynamics and Control* 32(5), 1569–1599.

Kawagoe, T. and Takizawa, H. (2008), ‘Equilibrium Refinement vs. Level-k Analysis: An Experimental Study of Cheap-Talk Games with Private Information’, *Games and Economic Behavior* 66, 238–255.

Kinderman, P., Dunbar, R. I. M. and Bentall, R. P. (1998), ‘Theory-of-Mind Deficits and Causal Attributions’, *British Journal of Psychology* 89, 191–204.

Kübler, D. and Weizsäcker, G. (2004), ‘Limited Depth of Reasoning and Failure of Cascade Formation in the Laboratory’, *Review of Economic Studies* 71, 425–441.

Mengel, F. (2009), ‘Learning Across Games’, Working paper, Instituto Valenciano de Investigaciones Económicas.

Nagel, R. (1995), ‘Unraveling in Guessing Games: An Experimental Study’, *American Economic Review* 85, 1313–1326.

Ohtsubo, Y. and Rapoport, A. (2006), ‘Depth of Reasoning in Strategic form Games’, *The Journal of Socio-Economics* 35, 31–47.

Palacios-Huerta, I. and Volji, O. (2009), ‘Field Centipedes’, *American Economic Review* 99(4), 1619–1635.

Penke, L., Denissen, J. J. A. and Miller, G. F. (2007), ‘The Evolutionary Genetics of Personality’, *European Journal of Personality* 21, 549–587.

Premack, D. and Wodruff, G. (1979), ‘Does the Chimpanzee have a Theory of Mind’, *Behavioral and Brain Sciences* 1, 515–526.

Robson, A. J. (2003), ‘The Evolution of Rationality and the Red Queen’, *Journal of Economic Theory* 111, 1–22.

Roth, G. and Dicke, U. (2005), ‘Evolution of the Brain and Intelligence’, *TRENDS in Cognitive Sciences* 9(5), 250–257.

Salmon, T. C. (2001), 'An Evaluation of Econometric Models of Adaptive Learning', *Econometrica* 69(6), 1597–1628.

Samuelson, L. (2001a), 'Analogies, Adaptation, and Anomalies', *Journal of Economic Theory* 97(2), 320–366.

Samuelson, L. (2001b), 'Introduction to the Evolution of Preferences', *Journal of Economic Theory* 97(2), 225–230.

Sandholm, W. H. (2010), *Population Games and Evolutionary Dynamics.* Book manuscript, to be published by MIT Press.

Sandholm, W. H. and Dokumaci, E. (2007), 'Dynamo: Phase Diagrams for Evolutionary Dynamics (Software suite)', http://www.ssc.wisc.edu/ whs/dynamo.

Sgroi, D. and Zizzo, D. J. (2009), 'Learning to play 3x3 games: Neural networks as bounded-rational players', *Journal of Economic Behavior and Organization* 69(1), 27–38.

Shapley, L. S. (1964), 'Some Topics in Two-Person Games', inM. Dresher, L. S. Shapley and T. A. W., eds, *Advances in Game Theory*, Princeton University Press, Princeton, New Jersey, pp. 1–28.

Stahl, D. O. (1993), 'Evolution of Smart$_n$ Players', *Games and Economic Behavior* 5, 604–617.

Stahl, D. O. (1999), 'Evidence based rules and learning in symmetric normal-form games', *International Journal of Game Theory* 28(1), 111–130.

Stahl, D. O. (2000), 'Rule Learning in Symmetric Normal-Form Games: Theory and Evidence', *Games and Economic Behavior* 32, 105–138.

Stahl, D. O. and Wilson, P. W. (1995), 'On Players'Models of Other Players: Theory and Experimental Evidence', *Games and Economic Behavior* 10, 218–254.

Steiner, J. and Stewart, C. (2008), 'Contagion through Learning', *Theoretical Economics* 3, 431–458.

Stennek, J. (2000), 'The Survival Value of Assuming Others to be Rational', *International Journal of Game Theory* 29, 147–163.

Vives, X. (1990), 'Nash Equilibrium with Strategic Complementarities', *Journal of Mathematical Economics* 19, 305–321.

Weibull, J.W. (1995), *Evolutionary Game Theory*, MIT Press, Cambridge Massachusetts.

Weissing, Franz, J. (1991), 'Evolutionary Stability and Dynamic Stability in a Class of Evolutionary Normal Form Games', in R. Selten, ed., *Game Equilibrium Models I. Evolution and Game Dynamics*, Springer-Verlag, pp. 29–97.

Wilcox, N. T. (2006), 'Theories of Learning in Games and Heterogeneity Bias', *Econometrica* 74(5), 1271–1292.

PAPER 4

Optimal Categorization

Erik Mohlin

Abstract. The importance of categorical reasoning in human cognition is well-established in psychology and cognitive science, and one of the most important functions of categorization is to facilitate prediction. This paper provides a model of optimal categorization. In the beginning of each period a subject observes a two-dimensional object in one dimension and wants to predict the object's value in the other dimension. The subject partitions the space of objects into categories. She has a data base of objects that were observed in both dimensions in the past. The subject determines what category the new object belongs to on the basis of observation of its first dimension. The subject predicts that its value in the second dimension will be equal to the average value among the past observations in the corresponding category. At the end of each period the second dimension is observed and the observation is stored in the data base. The main result is that the optimal number of categories is determined by a trade-off between (a) decreasing the size of categories in order to enhance category homogeneity, and (b) increasing the size of categories in order to enhance category sample size.

1. Introduction

Numerous psychological studies have demonstrated the importance of categorical reasoning in human cognition.[1] Categorical thinking also matters in many economic contexts: In financial markets, investors engage in "style investing" – the practice of allocating funds among classes of assets rather than to individual assets (Bernstein 1995, Barberis and Shleifer 2003). Rating agencies categorize firms in order to reflect the probability that a firm will default on its debt, an activity whose importance was highlighted by the recent financial crisis (Coval et al. 2009). In consumer markets, price discrimination has been extensively studied, but other forms of categorization

This paper has benefited from comments by Stefano Demichelis, Philippe Jehiel, Topi Miettinen, Paul Milgrom, Robert Östling, and Jörgen Weibull, as well as participants at presentations at the Third Nordic Workshop in Behavioral and Experimental Economics in Copenhagen, November, 2008, SUDSWEc in Uppsala, May 2009 and the Stockholm School of Economics. Financial support from the Jan Wallander and Tom Hedelius Foundation is gratefully acknowledged.

[1] For overviews of the voluminous literature see e.g. Laurence and Margolis (1999), or Murphy (2002).

also matter. Consumers categorize goods and services when deciding what to purchase, and this leads to segmentation of markets (Smith 1965). Firms may respond with marketing strategies that take advantage of the consumers' categorizations (Urban 1993, Punj and Moon 2002).[2]

In the psychological literature it is widely acknowledged that an important function of categories is to facilitate predictions about properties that are not immediately observable (Anderson 1990).[3] Prediction on the basis of categorical reasoning is relevant in situations where one has to predict the value of a variable on the basis of one's previous experience with similar situations, but where the past experience does not include any situation that was identical to the present situation in all relevant aspects.[4] In such situations one must consider both one's experience of what happened in previous situations, and how similar those situations were to the present situation. This can be done by dividing the experienced situations into categories, such that situations in the same category are similar to each other. When a new situation is encountered one determines what category this situation belongs to, and the past experiences in this category are used to make a prediction about the current situation. Two stylized features of this process need to be stressed – and are incorporated into the formal model below: First, predictions about a particular category are generally formed only on the basis of objects that were put into that category in the past, not on the basis of objects that were put into other categories (Malt et al. 1995, and Murphy and Ross 1994). Second, a prediction about a particular object is generally based only on what category the object belongs to, and does not take into account within-category correlations between properties. This means that roughly the same prediction is made for all objects in the same category (Krueger and Clement 1994).[5]

Despite the importance of categorical reasoning, there are only a few explicit models of categorization in economics. Moreover, the question of optimality has rarely

[2] From a choice-theoretic perspective it can be noted that categorical reasoning generates insensitivity to differences and hence a categorizing subject might exhibit intransitive preferences.

[3] Related to this there are also studies of inductive inference on the basis of categories (e.g. Rips 1975 and Osherson et al. 1990).

[4] I will use the term 'prediction' both for the case of predicting a stochastic variable that is not yet realized and for the case of assessing the value of an unobserved non-stochastic variable (or an unobserved realization of a stochastic variable).

[5] Krueger and Clement (1994) find that predictions about temperature on different dates of the year vary discontinuously between months, so that the temperature on the last day of January is underestimated (too cold) and the temperature on the first of February is overestimated (too warm). This is naturally interpreted as the result of using months as categories for making predictions. I want to point out that in order for this explanation to hold one must assume that predictions are insensitive to within category (month) trends – in line with the second stylized feature mentioned above. Otherwise one would be able to predict the temperature at the end of January correctly, even without using data from the month of February.

been discussed, and those who have done so, e.g. Fryer and Jackson (2008), assume an exogenous number of categories. In this paper I ask which categorizations that are optimal in the sense that they minimize prediction error – a notion that is made precise below. In particular, I wish to derive the optimal number of categories without imposing any exogenous costs and benefits of the number of categories. Instead *both costs and benefits are derived endogenously from the objective of making accurate predictions*. The advantage of fine grained categorizations is that objects in a category are similar to each other. The advantage of coarse categorizations is that a prediction about a category is based on a large number of observations. Comparative statics reveal how the optimal categorization depends on the number of observations, as well as on the frequency of objects with different properties. To the best of my knowledge this is the first paper to investigate categorizations that are optimal from the point of view of prediction.[6] Related literature, including Fryer and Jackson (2008), Al-Najjar and Pai (2009) and Peski (2007) is discussed below.

Given that we reason in terms of categories, why should we be interested in *optimal* categorizations? From an evolutionary perspective we would expect humans to have developed categories, and categorization procedures, that tend to result in categorizations that generate predictions that induce behavior that maximize fitness. It seems reasonable to assume that fitness is generally increasing in how accurate the predictions are. For instance, a subject encountering a poisonous plant will presumably be better off if she predicts that the plant is indeed poisonous, rather than nutritious. For this reason we would expect to find that humans employ categorizations that are at least approximately optimal, in the sense that they minimize prediction error. Note that the set-up does not presume the existence of any natural kinds (Quine 1969). There does not have to exist an objectively true categorization "out there". The optimal categorization is a framework we impose on our environment in order to predict it.

The dominant view within psychology is that the number of categories (the coarseness of the categorization) is determined by another trade-off (Medin 1983). Like in this paper, the benefit of small categories is supposed to be within-category homogeneity of objects. But, unlike this paper, the benefit of having a fewer larger categories is supposed to be that one needs to observe fewer properties of an object in order to categorize it as belonging to a large category. A virtue of the explanation put forward in this paper is that it connects a main purpose of categorization, namely prediction, both with the value of many small categories and with the value of a few large categories.

[6] After this paper was written it came to my attention that Al-Najjar and Pai (2009) discuss similar issues. However, it should be noted that the first version of Al-Najjar and Pai's paper is dated December 2008, whereas the first version of this paper was presented publicly on November 14, 2008, at the Third Nordic Workshop in Behavioral and Experimental Economics in Copenhagen.

The model is centered around a subject who lives for a certain number of periods. First she goes through a learning phase and then a prediction phase. In each period of the *learning phase* she observes an object, represented by a pair of numbers (x, y). All objects are independently drawn from the same distribution, and are stored in a data base. At the beginning of the learning phase the subject is endowed with a categorization, which is kept fixed for the subject's whole life time. A categorization is a set of categories which together partition the set of objects. Each object's category membership is determined by its x-value. In the beginning of each period of the *prediction phase* the subject encounters a new object and observes the x-value but not the y-value. The y-value has to be predicted with the help of the object's x-value and the data base of past experiences. The new object is put in one of the categories on the basis of its x-value. The empirical mean y-value, of the previously experienced objects in that category, serves as prediction for the y-value of the new object. At the end of the period, after the prediction has been made, the y-value is revealed and the information is added to the data base.[7]

To fix ideas, think of a physician who encounters a new patient in each period. The x-value could represent information about a patient's personal characteristics such as weight, blood-pressure, or aspects of the patient's medical history. The y-value could represent some dimension of the patient's future health. During the learning phase the physician goes to medical school and learns a set of categories while observing various patients' characteristics together with their subsequent health state. In the prediction phase she works in a hospital: In the beginning of each period she receives information about a patient's personal characteristics, and has to make a prediction about some aspect of the patient's health. In order to make such a prediction she assigns the new patient to a category and predicts that the outcome for this patient will be like the empirical average outcome among previous patients in that category. At the end of each period she can observes the outcome for the current patient.

Alternatively, think of color concepts. The subset of the spectrum of electromagnetic radiation that is visible to the human eye allows for infinitely fine grained distinctions. More precisely all possible colors can be described as points in the space of the three dimensions hue, saturation and lightness.[8] However, in every day reasoning and discourse we seem to employ only a coarse color classification, using words such as red, green, turquoise, etcetera. We could have sliced up the space of colors differently. Indeed there are other cultures where even the basic color categories are different from

[7] The assumption that all y-values are eventually observed is a simplification that may not hold in applications. Sometimes y-values will only be observed if certain actions are taken with respect to the object and these actions may depend on the prediction that has been made.

[8] For an introduction see http://en.wikipedia.org/wiki/HSL_and_HSV.

the ones used by speakers of the English language (though the relativity seems to follow certain principles; see e.g. Kay and Maffi 1999 and references therein). Presumably the color categorizations that were developed and passed on to new generations, were successful in the kind of environments that we faced.

Prediction error is measured as the squared difference between the prediction and the actual y-value of the object. Using the probability density function over the set of objects one can define the *expected prediction error* of a categorization. Expectation is taken over the set of data bases that the subject may encounter. The expected prediction error is minimized by an *optimal categorization*. This is the relevant notion of optimality for the many categories that are learned early in life through socialization and education. From an evolutionary perspective we expect humans to have developed, and to pass on, categorizations that minimize prediction error in the relevant environments. The reader might find it useful to think of this in terms of a principal-agent framework. Evolution can then be represented by the principal, and humanity is represented by the agent. The principal knows the distribution of objects and computes the optimal categorization and gives it to the agent. The agent does not know the distribution of objects so she uses the categorization from the principal to make predictions.

In other cases we develop new categories only *after* having accumulated a data base. In this case an evolutionary perspective implies that we should expect to find that humans employ categorization *procedures* that result in *categorizations* that are at least approximately optimal, in the sense that they minimize prediction error. An alternative notion of optimality for this case is discussed in section 3.3.

The main result of this paper is that the optimal number of categories is determined by a trade-off between the value of within-category similarity of objects and the value of having many stored observations in each category. Increasing the number of categories has two effects. (a) The average size of each category decreases and thus the differences between objects that belong to the same category will be relatively small. (b) The average number of experienced objects in each category decreases. Thus generalizations about a category are based on a smaller sample, making inferences from observed objects to future cases less reliable. Note that this trade-off does *not* depend on any exogenous cost of categories. The trade-off sheds light on the phenomenon of basic-level categories, which has received much attention from psychologists; the most salient level of categorization is neither the most fine-grained, nor the most general level of categorization (Rosch et al. 1976). The model can also explain why experts tend to have a more fine grained conceptual structure than laymen (Tanaka and Taylor 1991, Johnson and Mervis 1998). Furthermore, comparative statics with respect to the distribution of objects with different properties show that (i) the larger the variability in the y-dimension, the larger is the optimal number of categories, and (ii) the more frequent objects in one subset of the x-dimension are, the larger is the optimal number of categories in that subset. In particular, assuming that the relationship between x- and y-values is given by a linear regression model, the optimal number of categories is decreasing in the variance of the error term and increasing in the slope of the regression line. Finally some extensions of the model are discussed: The possibility of choosing in what category to make observations is investigated and related to the interplay of observation and concept formation in science.

It should be emphasized that the inference, from properties of objects in the data base, to the unobserved property of the present object, is *not* Bayesian. In particular, the subject does not have a prior about an object's properties before it is categorized. On the contrary, the model of this paper is intended to shed some light on how priors are generated. When an object is categorized, the data base is used to form a point prediction about the new object, in a non-Bayesian, frequentist way. This point prediction should be interpreted as a prior (point) belief. One could also think of a more complex model where the data base is used to form probabilistic beliefs, i.e. the prediction for objects in a certain category takes the form of a density over Y. Giloba et al. (2008) argue that Bayesian decision theory needs to be complemented with a theory of belief formation that accounts for how priors are formed. They argue that the vast majority of decision problems are such that there is too little information to adopt a prior that is based on inferences from cases that are identical (in relevant dimensions) to the present case, and yet there is too much information to apply a symmetric prior based on the principle of insufficient reason. Related to this, Binmore (2007) argues that the Bayesian approach is inappropriate in "large worlds", i.e. decision problems involving large state spaces, and claims that we need a theory of belief formation in such settings. It is hoped that this paper may provide a step towards such a theory. (For a discussion of these matters see also Morris (1995).)

Categories are closely related to concepts. Categories can be said to be defined by concepts in the sense that an object belongs to a category if an only if it falls under the corresponding concept. Conversely, categorization is one of the most important functions of concepts (see Solomon et al. 1999 about other functions). One might suggest that we use categories because language is categorical and say that a categorization is optimal if it is induced by a language that is optimal in some sense. Language is undoubtedly important in shaping our concepts and categories, but concepts came prior to language in evolution – there are animals that use concepts even though they do not use language – and children can use certain concepts before they have a language.[9] Therefore I suggest that we try to explain the use of categories without reference to language. In addition to this, a language usually allows many different categorizations of the same subject matter. Therefore the optimal categorization is under-determined by the demands of communication.

The rest of the paper is organized as follows. Section 2 describes the model and defines prediction error and optimality. The results are developed in section 3, and discussed in section 4. Section 5 reviews related literature, and section 6 concludes. All proofs are in the appendix, section 6.

2. Model

2.1. Subject and Objects. A subject lives for T periods; first a learning phase of $L < T$ periods, and then a prediction phase of $T - L$ periods. In each period $t \in \{1, ..., T\}$ she encounters an object, which is represented by a point $v_t = (x_t, y_t)$ in a two-dimensional Euclidean space $V = X \times Y$. The set X is a closed interval $[a, b]$ on the real line, and the set Y can be any interval on the real line. All objects are drawn independently according to a continuous probability density function $f : V \to [0, 1]$, satisfying $f(v) > 0$ for all $v \in V$.[10] All experienced objects are stored in a data base, so at the beginning of any period $t > 1$ the subject has a data base $v^{t-1} = (v_1, ..., v_{t-1}) \in V^{t-1}$. In each period $t \in \{1, ..., L\}$ of the learning phase the subject observes each object in both dimensions. In the beginning of each period $t \in \{L + 1, ..., T\}$ of the prediction phase she observes the x-value, x_t, of an object v_t, and not its y-value, y_t. She makes a prediction about y_t on the basis of x_t, and the data base v^{t-1}. At the end of the period uncertainty is resolved; the subject observes y_t, and updates the data base. Thus learning does not only occur in the learning phase but continues through the whole life time.

The set-up described here can be extended to an $n = n_x + n_y$ dimensional Euclidean space $V = X \times Y$, where the subspace $X \subseteq \mathbb{R}^{n_x}$ is compact and convex. For simplicity

[9] Regarding animals there is evidence that pigeons have concepts, at least in a way that enables them to categorize objects (Herrnstein 1979, Herrnstein et al. 1976). There are also studies indicating that rhesus monkeys (Hauser 1996) and cotton-top tamarins (Uller 1997) have simple numerical concepts. Regarding children Franklin et al. (2005) provide evidence that toddlers have a pre-linguistic understanding of color concepts.

[10] By letting $Var(y|x) = 0$ for all $x \in X$, the model can accommodate the special case of a deterministic relationship between X and Y. This would describe a situation where the subject knows all the factors that influence the y-value, except for the factor that is represented by X.

I develop my results for the case of one observable and one unobservable dimension; $n_x = n_y = 1$.[11]

2.2. Categories. At the beginning of period 1 the subject is endowed with a categorization that is fixed for the rest of the subject's life. A category C_i is a subset of V. A categorization is a finite set of categories $C = \{C_1, ..., C_k\}$ that constitutes a partitioning of V. Let X_i be the projection of C_i onto X. Since the category membership of an object only depends on the object's x-value, the collection of sets $\{X_1, ..., X_k\}$ form a partitioning of X, and we can write $C_i = X_i \times Y$. Each set X_i is assumed to be the union of finitely many intervals.[12] The relative size of categories is constrained by some (small) number $\rho \in (0, 1)$ such that $\Pr(x \in X_i) / \Pr(x \in X_j) > \rho$ for all i and j. For the case of a finite number of categories this implies that all categories have positive probability. (When the number of categories goes to infinity the assumption implies that no category becomes relatively infinitely larger than another category.) The set of feasible categorizations is denoted Ψ.[13]

It might seem problematic to assume that categories in the same categorization are mutually exclusive, since we have many categories that are not mutually exclusive. This is the case for hierarchically organized concepts such as the two categories of stone and granite. However, we generally do not use such overlapping categories for the same prediction tasks. If I am interested in whether an object will burn when thrown on the fire I might categorize the object as made of stone rather than wood, and infer that it will not burn. In this context it is useless to know whether the object is of granite or not. But if I want to build a house it may be useful to employ a narrower categorization of materials, since granite is more solid than e.g. limestone.

In section 4.5 I investigate what happens when one allows for categorizations that are not jointly exhaustive (but mutually exclusive).

2.3. Prediction . For each category $C_i \in C$, and for date t, the subject has a prediction $\hat{y}_{it}$ about the y-value of objects in that category. As discussed above, it will

[11] Although it would require more modifications, it should also be possible to generalize the results to the case of a set Y that is not an interval. For instance, if one wants to predict the probability of a stock market crash in a particular country it would be natural to let $Y = \{0, 1\}$.

[12] If categories are only composed of one interval the categories are required to be convex. Gärdenfors (2000) argues that we should expect the extension of natural concepts to be convex on the grounds that convex concepts are easier to learn than non-convex concepts. Warglien and Gärdenfors (2008) argue that it will be easier for communicating parties to agree on a joint meaning if the concepts are convex. Thus a restriction to convex categories might be natural for certain kinds of concepts but I will work with the more general assumption that the categories are the union of finitely many intervals.

[13] The model allows a for many different sets to constitute categories. There is evidence that human categorization is indeed characterized by such flexibility (Ashby and Waldron 1999, McKinley and Nosofsky 1995).

be assumed that the prediction equals the mean of all previously experienced objects in that category. If the data base for a certain category is empty then the prediction for that category is equal to the mean of all previously encountered objects. Let

$$D_{it} = \{s \in \mathbb{N} : s < t \wedge v_s \in C_i\}.$$

This is the set of dates, prior to date t, at which objects in category C_i were observed. Let $m_{it} = |D_{it}|$, so that $\sum_{i=1}^{k} m_{it} = t - 1$, for all t. Thus at date $t > L$ the prediction for category i is

$$\hat{y}_{it} = \begin{cases} \frac{1}{m_{it}} \sum_{s \in D_{it}} y_s & \text{if } m_{it} > 0 \\ \hat{y}_t & \text{if } m_{it} = 0 \end{cases},$$

where

$$\hat{y}_t = \frac{1}{t-1} \sum_{s=1}^{t-1} y_s.$$

This means that when the data base does not contain any objects in the category that object v_t belongs to, then the prediction for this object is made on the basis of all objects currently in the data base. This seems like a natural assumption, but one could make other assumptions, and this would not affect the results of the paper, except proposition 8. The reason is that most results concern the case of a large number of observations relative to the number of categories, so that the probability of an empty category is negligible.

One can also modify the model more radically and assume that the subject always has at least one object in each category. Formally this can be done by assuming that the agent is endowed with these observations together with the categories, in period 1. Again, almost all the results will go through under this alternative assumption, the only exception being proposition 8.

2.4. Prediction Error and Optimality. For any object v_t that the subject may encounter there is a unique category C_i such that $v_t \in C_i$. For any data base $v^{t-1} \in V^{t-1}$ that the subject may have at date t the prediction y_{it} is then determined according to the definition above. The prediction error is measured as the squared Euclidean distance between the predicted value $\hat{y}_{it}$ and the true value y_t:

DEFINITION 10. *For any data base v^{t-1} and any new object $v_t \in C_i$ the prediction error is*

$$PE\left(C, v_t, v^{t-1}\right) = \left(y_t - \hat{y}_{it}\right)^2.$$

At time t the (unconditional) expected prediction error of categorization C is

$$EPE\left(C, t\right) = \mathbb{E}\left[PE\left(C, v_t, v^{t-1}\right)\right].$$

Here expectation is taken over objects in V and over data bases in V^{t-1}. Summing over the $T-L$ prediction tasks that the subject has to perform, one can define the total expected prediction error of a categorization.

DEFINITION 11. *The total expected prediction error $EPE(C,T,L)$ of a categorization C is*

$$EPE(C,T,L) = \frac{1}{T-L} \sum_{t=L+1}^{T} EPE(C,t)$$

This is used to define the notion of an optimal categorization:

DEFINITION 12. *An optimal categorization is a categorization $C \in \Psi$ that minimizes $EPE(C,T,L)$.*

3. Results

3.1. Preliminary Results . In order to illuminate the basic trade-off I will investigate the expected prediction error conditional on a data base v^{t-1}, defined as

$$EPE\left(C, v^{t-1}\right) = \mathbb{E}\left[PE\left(C, v_t, v^{t-1}\right) | v^{t-1}\right].$$

Note that

$$\Pr\left((x,y) \in C_i\right) = \Pr\left(x \in X_i\right) = \int_{x \in X_i} \int_{y \in Y} f(x,y)\, dx dy,$$

and define

$$f(y|x \in X_i) = \frac{1}{\Pr(x \in X_i)} \int_{x \in X_i} f(x,y)\, dx.$$

Also define

$$Var(y_i) = Var(y|x \in X_i).$$

Using this one can show.

LEMMA 8. *The expected prediction error for a categorization C, conditional on a data base v^{t-1}, is*

$$EPE\left(C, v^{t-1}\right) = \sum_{i=1}^{k} \Pr\left(x \in X_i\right)\left(Var(y_i) + \left(\hat{y}_{it} - \mu_i\right)^2\right).$$

This expression reveals the basic trade-off that determines the optimal number of categories. The term $Var(y_i)$ measures how similar one can expect objects in the same category to be with respect to distance in the y-dimension. The term $\left(\hat{y}_{it} - \mu_i\right)^2$ measures how close predictions are to the actual averages of the corresponding categories. The optimal categorization strikes a balance between the goal of having a low within category variance and the goal of estimating the category mean correctly.

Fix the date t and take expectation of $EPE\left(C, v^{t-1}\right)$ with respect to the data bases of size $t-1$. Then one obtains:

LEMMA 9. *The expected prediction error for a categorization C, at time t, is*

$$EPE\left(C,t\right)=\sum_{i=1}^{k}\Pr\left(x\in X_i\right)Var\left(y_i\right)\left(1+\sum_{r=1}^{t-1}\Pr\left(m_{it}=r\right)\frac{1}{r}\right)$$
$$+\sum_{i=1}^{k}\Pr\left(x\in X_i\right)\Pr\left(m_{it}=0\right)\mathbb{E}\left[\left(\hat{y}_t-\mu_i\right)^2|m_{it}=0\right],$$

where m_{it} has a binomial distribution

$$\Pr\left(m_{it}=r\right)=\frac{(t-1)!}{r!\,(t-1-r)!}\left(\Pr\left(x\in X_i\right)\right)^r\left(1-\Pr\left(x\in X_i\right)\right)^{t-1-r}.$$

It is difficult to derive general results about the optimal categorization unless L or T is large (though finite). The difficulties partly stem from the binomial expression in $EPE\left(C,t\right)$ and partly from the fact that no assumptions are made about f except that it is continuous. Therefore most of the results presented below will make the assumption that L, or T are sufficiently large. This should not be a problem, because during the process of learning concepts and categories as a child or student, a subject accumulates a large data base. The categorization abilities of the adult subject should then indeed be captured by results proved for the case of a large but finite data base.

It will be fruitful to decompose the within-category variance in the y-dimension, $Var\left(y_i\right)$, into the contribution of the within-category average conditional variance

$$\mathbb{E}\left[Var\left(y|x\right)|x\in X_i\right]=\int_{x\in X_i}\frac{f\left(x\right)}{\Pr\left(x\in X_i\right)}Var\left(y|x\right)dx,$$

and the within-category variance of the conditional expected value

$$Var\left(\mathbb{E}\left[y|x\right]|x\in X_i\right)=\int_{x\in X_i}\frac{f\left(x\right)}{\Pr\left(x\in X_i\right)}\left(\mathbb{E}\left[y|x\right]-\int_{x\in X_i}\frac{f\left(x\right)}{\Pr\left(x\in X_i\right)}\mathbb{E}\left[y|x\right]dx\right)^2dx.$$

The following lemma establishes their connection:

LEMMA 10. *The within-category variance is the sum of the within-category average conditional variance, and the within-category variance of the conditional expected value;*

$$Var\left(y_i\right)=\mathbb{E}\left[Var\left(y|x\right)|x\in X_i\right]+Var\left(\mathbb{E}\left[y|x\right]|x\in X_i\right).$$

Before providing results regarding optimal categorizations, I establish that such categorizations exist:

LEMMA 11. *Suppose that the number of categories and the number of unconnected subsets of each category are uniformly bounded above. Then, for any t, v^{t-1}, and*

$L < T$, there exist solutions to the problems of minimizing $EPE\,(C, v^{t-1})$, $EPE\,(C, t)$, or $EPE\,(C, T, L)$, with respect to C.

It can be noted that there is no guarantee that any of these solutions are unique, thus allowing for a (mild) form of conceptual relativism.

3.2. Properties of Optimal Categorizations . The following proposition establishes that if the learning phase is sufficiently long in relation to the prediction phase then the optimal number of categories is less than the number of observations made during the observation phase. Also, for any given length of the learning phase, if the prediction phase is sufficiently long, then the optimal number of categories is less than the number of observations made during the life time. Finally, as the number of observations goes to infinity, the average number of object in each category goes to infinity.

PROPOSITION 1. ***(a)*** *There are finite L' and T', with $L' < T'$, such that if $L' < L < T < T'$, then any optimal categorization for L and T, has $k < L$.* ***(b)*** *For any finite L there is a finite $T' > L$ such that if $T > T'$ then any optimal categorization for L and T, has $k < T$.* ***(c)*** *For any finite L, if $T \to \infty$ then the optimally $k \to \infty$ and $k/T \to 0$.*

The following is an immediate consequence of the fact that it is never optimal to have more categories than objects:

COROLLARY 1. *If t is large enough then there exists a solution to the problems of minimizing $EPE\,(C, v^{t-1})$, or $EPE\,(C, t)$, even when one does not assume a uniform bound on the number of categories. Also, if L is large enough then the same holds for minimization of $EPE\,(C, T, L)$.*

The next proposition says that if the learning phase, or the prediction phase, is sufficiently long, then the optimal categorization has more than one category, provided that the conditional mean $E\,(y|x)$ is not constant over X.

PROPOSITION 2. *If f is such that there are two disjoint and mutually exclusive sets $E, F \subset X$, with $\mathbb{E}\,(y|x \in E) \neq \mathbb{E}\,(y|x \in F)$, then there are finite L' and T' such that if $L > L'$ or $T > T'$ then any optimal categorization has $k > 1$.*

Propositions 1 and 2 together provide an explanation for why we typically employ categorizations that are neither maximally fine grained – with one object in each category – nor maximally general – with all object in the same category. This is discussed further section 4.2.

In the special case when the conditional distribution is constant over X, it is optimal to have only one category, regardless of the size of the data base:

PROPOSITION 3. *If $f(y|x) = f(y|x')$ for all $x, x' \in X$ then any optimal categorization has $k = 1$.*

Now consider two subjects 1 and 2, with different learning phase L and different total number of observations T. Assume $T_2 > T_1$ and $L_2 > L_1$. (In section 4.3 subjects 1 and 2 are interpreted as being a layman and an expert, respectively.) The model predicts that if the differences between the two subjects are large enough then it is optimal for the individuals to have $k_2 > k_1$.

PROPOSITION 4. ***(a)*** *There are finite L' and T', with $L' < T'$, such that if $L_2 - L_1 > L'$ and $T_2 - T_1 > T'$, then optimally $k_2 > k_1$.* ***(b)*** *For any finite difference $L_2 - L_1$ there is a finite T' such that if $T_2 - T_1 > T'$ then optimally $k_2 > k_1$.*

The next three propositions concern the relationship between the density $f(x, y)$ and the optimal categorization. The first result considers the marginal density over X, i.e. $f(x)$. The more common objects from one subset of X are, the more fine-grained should the optimal categorization for that subset be:

PROPOSITION 5. *Consider a proper subset $E \subseteq X$, and two densities f_0 and f_1, such that $f_0(y|x) = f_1(y|x)$ for all $x \in E$. Suppose there is some $\alpha > 1$ such that $\alpha f_0(x) = f_1(x)$ for all $x \in E$. Then the lowest optimal number of categories in E is at least as large with f_1 as with f_0.*

This is a generalization of the result in Fryer and Jackson (2008), that less frequent objects will be categorized more coarsely. Their result assumes a fixed number of categories, whereas mine does not. They relate the result to the possibility that ethnic minorities will be categorized more coarsely than majorities. This will tend to lead to more stereotypical predictions about the minority than the majority.

The next result concerns the effect of the conditional variance, $Var(y|x)$, on the optimal categorization.

PROPOSITION 6. *Consider two densities f_0 and f_1, such that $f_0(x) = f_1(x)$, $\mathbb{E}_{f_0}[y|x] = \mathbb{E}_{f_1}[y|x]$ and $Var_{f_1}(y|x) > Var_{f_0}(y|x)$ for all $x \in X$. There are L' and T' such that if $L > L'$ or $T > T'$, then the lowest optimal number of categories is at least as large with f_1 as with f_0.*

We saw above that

$$Var(y_i) = \mathbb{E}[Var(y|x)|x \in X_i] + Var(\mathbb{E}[y|x]|x \in X_i).$$

Proposition 6 concerns comparative statics with respect to the first term on the right hand side. Comparative statics with respect to the second term on the right hand side requires more detailed assumptions about the distribution f. For this reason I now restrict attention to the following special case: Suppose $X = [0,1]$ and $Y = \mathbb{R}$ and suppose that the relation between them is described by the classical linear regression model;

$$y = \alpha + \beta x + z,$$

where $z \sim N(0, \sigma^2)$. Furthermore assume that x is uniformly distributed on X. Assume also that the subject only makes one prediction during her life, i.e. $T - L = 1$ (extension to $T - L > 1$ is straightforward but does not add insight). Finally, for simplicity, also assume that subjects are endowed with one observation in each category already in period 1, as mentioned in section 2.3. (The results become more tractable with this assumption but the general insight is unaltered.) Under these assumptions we have the following result:

PROPOSITION 7. ***(a)*** *For any T and L the number of categories in the optimal categorization is unique and all categories have the same length along the x-axis: If the optimal number of categories is k, then the optimal categories satisfy $X_i = [a_i, b_i)$, and $b_i - a_i = 1/k$ for all $i < k$. The k^{th} category satisfies $X_k = [b_{k-1}, 1]$, and $1 - b_{k-1} = 1/k$.* ***(b)*** *The optimal number of categories is increasing in β and decreasing in σ^2.*

Recall that for the linear regression model it holds that

$$\beta = \frac{Cov(x, y)}{Var(y)},$$

so increasing covariance of x and y increases the optimal number of categories. Increasing the conditional variance of y decreases the optimal number of categories. This result is very intuitive: If the covariance is large then the categories have to be narrow in order to keep the heterogeneity of objects in each category within limits. If the variance of y is large then (in line with proposition 6) the categories have to be broad in order to contain enough objects to allow reasonably accurate estimates of the means of each category.

3.3. Categorization Conditional on a Data Base . An evolutionary perspective suggests that, we should expect that humans have developed, and pass on, categorizations that minimize prediction error for the kind of data bases that one can expect to encounter in life. Similarly one would expect that over time a profession has developed a categorization that minimize prediction error for the kind of data bases that members of that profession tend to encounter. The definition of an optimal categorization, as minimizing *unconditional* expected prediction error, is intended to capture this evolutionary adaptation.

However, as mentioned in the introduction, it also happens that new concepts are developed *after* a data base has been accumulated – e.g. for some area of investigation where one did not have useful concepts before. An evolutionary perspective then suggests that humans might have developed adaptive processes for categorization of given data bases. Such processes would tend to result in categorizations that minimize expected prediction error *conditional* on the data base. The expected prediction error conditional on a data base, $EPE\left(C, v^{t-1}\right)$, was defined above. It can be used to define a second notion of optimality.

Definition 13. *The optimal categorization conditional on a data base v^{t-1}, is the categorization $C \in \Psi$ that minimizes $EPE\left(C, v^{t-1}\right)$.*

An important question is of course what algorithms or rules of thumb that a subject will use to determine what the optimal categorization conditional on a data base v^{t-1}, is. A categorizing subject should not be assumed to know the distribution f, because if the subject did know f, then there would be no need to base predictions on categorization, rather than using knowledge of the density f directly. Still, we would expect the rules of thumb to yield approximately optimal categorizations. Thus, as analysts we might be willing to assume that subjects act *as if* they optimized on the basis of knowledge of f. Then we can predict that subjects will use categories that are optimal in the sense of definition 13.

A more explicit, and more realistic, model would assume that the subject follows some other rule than directly choosing a category that minimizes $EPE\left(C, v^{t-1}\right)$. One rule of thumb that seems reasonable is to choose a categorization C that minimizes some estimator of $EPE\left(C, v^{t-1}\right)$. For example, one could use the following estimator:

Definition 14. *The sample prediction error for a categorization C, with nonempty categories, conditional on a data base v^{t-1}, is*

$$SPE\left(C, v^{t-1}\right) = \sum_{i=1}^{k} \frac{m_{it}}{t-1} \widehat{Var\left(y_i\right)} + \frac{1}{k-1} \sum_{j=1}^{k} \left(\hat{y}_j - \hat{y}\right)^2,$$

where

$$\widehat{Var\left(y_i\right)} = \frac{1}{m_{it}-1} \sum_{s \in D_{it}} \left(y_s - \hat{y}_{it}\right)^2.$$

The first term of $SPE\left(C, v^{t-1}\right)$ contains two factors. The factor $m_{it}/\left(t-1\right)$ is an estimator of $\Pr\left(x \in X_i\right)$, and the factor $\widehat{Var\left(y_i\right)}$ is an estimator of $Var\left(y_i\right)$. The second term of $SPE(C, v^{t-1})$ is intended to be an estimator of the factor

$$\sum_{i=1}^{k} \Pr(x \in X_i)(\hat{y}_{it} - \mu_i)^2,$$

in $EPE(C, v^{t-1})$. It is beyond the scope of this paper to explore the properties of this rule for creating categories given a data base, but it is clear that it involves a trade off between making categories small in order to decrease the first term, and making categories large in order to decrease the second term.

4. Discussion

4.1. Why use Categories? This paper has taken the fact as given that we use categories, and then characterized what categorizations that are optimal for the purpose of minimizing prediction error. One might ask if it is optimal to use categorizations at all, rather than some other estimation procedure, rather than say kernel-based estimation. This is undoubtedly an important question but I will only make a few remarks on why evolution might have favored the use of categorizations. One advantage of basing predictions of categories is that at each point in time one only needs to keep track of as many predictions as there are categories. If one uses e.g. kernel-based estimation instead, then one has to compute a distinct prediction for each object that one might encounter. This also means that after each observation all predictions have to be updated. In contrast, if one uses categories, then only the predictions associated with one category need to be updated after each observation.

4.2. The Optimal Number of Categories and the Basic Level . In studies of concepts and categorization with hierarchically organized concepts (e.g. animal – bird – robin) it is found that there is a privileged level in the hierarchy, called the basic level. Generally this level is named spontaneously in categorization tasks, learned first by children, and is in other ways salient (Rosch et al. 1976). The basic level is neither the most general level nor the most detailed level (e.g. bird rather than the superordinate category animal or the subordinate category robin). If categories are useful because they facilitate prediction then it might seem that it should be optimal to have an infinitely fine grained conceptual structure, since the narrower a category is, the more precise are the predictions that can be inferred from category membership. For example, if something is categorized as a bird then one can infer that it lays eggs and has wings, but if it is categorized (more finely) as a penguin, then one can also infer that it cannot fly but can dive. Since we do not use infinitely fine-grained category structures there must be some other factor that decreases the benefit of narrow categories. I have argued that this factor is constituted by the need to have a sufficiently large sample in each category to generalize from.

The dominant view in psychology has instead been that the cost of fine grained categorizations has to do with the difficulty of categorizing objects into fine grained categories: In order to make a more fine-grained categorization one has to observe more properties of an object.[14] A virtue of the explanation put forward in this paper is that it connects a main purpose of categorization, namely prediction, both with the value of many small categories and with the value of a few large categories.[15] In the end it is of course an empirical question as to which theory is the best. However, it is difficult to come up with a clean test. The reason is that lower level categories both contain less objects and are associated with more stringent conditions for application. Experimentally one could try to find a superordinate category and a subordinate category which are equally easy to apply. The conventional psychological explanation would then predict that the basic level will not be the superordinate of these two categories. In contrast, my explanation would predict the superordinate category to be basic if the subordinate category contains too few exemplars, or is associated with too much variance. Of course the explanations could be viewed as complementary. Both may describe forces that shape our categorizations.

4.3. Experts and Laymen . Experts tend to have a more fine grained conceptual structure than laymen (Tanaka and Taylor 1991, Johnson and Mervis 1998). This can

[14] Rosch and Mervis (1975) suggest that the basic level is the level that strikes an optimal balance between maximizing within category similarity and minimizing between-category similarity. Rosch et al. (1976) define cue validity as the conditional probability that an object belongs to a certain category, given that it has a certain feature (cue). This measures how easy it is to categorize an object as belonging to the category. Cue validity is always maximized for the highest, most inclusive, level in the hierarchy; Murphy (1982). medin (1983) defines category validity as the conditional probability that an object has a feature given that it belongs to a certain category. This measures the predictive power of categories, and is maximized for the most specific categories, at the bottom of the hierarchy. He suggests that the basic level represents the optimal trade off between cue validity and category validity, i.e. between easy classification and sharp predictions. Jones (1983) formalized this suggestion as the maximization of the product of cue validity and category validity. Corter and Gluck (1992) incorporate the utility of communication into this approach. A related explanation, closer to the "theory theory" of concepts, is provided by Markman and Wisniewski (1997).

[15] Recently, Porthos and Chater (2002) put forward an account of categorization that builds on the idea that the simplest categorization will be preferred. Simplicity is identified with code length. In their model there is a trade-off between reducing the number of categories (thereby simplifying the representation of similarity of categories) and reducing the number of objects within each category (thereby simplifying the representation of the within-category similarity). Thus on this account the optimal categorization is one that maximizes simplicity, rather than predictive success. There are some results establishing a link between simplicity and prediction (see Chater 1999). Another link between prediction and simplicity is explored by Gilboa and Samuelson (2008). Their approach is connected with the present paper since not using too many categories can be a way of keeping a theory simple.

be explained in the present model, with the help of proposition 4. Consider a layman with a learning phase of length L_1 and a prediction phase of $T_1 - L_2$ periods. Suppose the optimal number of categories for this person is k_1. An expert is distinguished by that she goes through more extensive training, L_2, or a longer prediction phase $T_2 - L_1$, than the layman The model predicts that if these differences between an expert and a layman, then it is optimal for the expert to have larger number of categories than the layman; $k_2 > k_1$.

This may also explain why some populations use a more fine-grained category structure than other populations: For instance, people in traditional subsistence cultures tend to have more specific biological categories than e.g. American college students (Berlin et al. 1973, Solomon et al. 1999). Of course there are other possible explanations for this phenomenon.

4.4. Interplay of Observation and Categorization . Consider a scientist who wants both to develop new theories, in the form of new concepts, and to make new observations, in order to make better predictions. The scientist can influence what observations she makes by choosing to perform some experiments rather than others. For a given categorization the above model can be modified to allow the subject to choose in what categories to make her observations, with the purpose of minimizing the expected prediction error $EPE\left(C, v^{t-1}\right)$. Thus the subject chooses the numbers of observations in different categories $\{r_i\}_{i=1}^k$, such that $\sum_{i=1}^k r_i = t-1$, in order to minimize

$$\begin{aligned}
&\mathbb{E}\left[EPE\left(C, v^{t-1}\right)|\{m_{it}\}_{i=1}^k = \{r_i\}_{i=1}^k\right] \\
&= \sum_{i=1}^k \Pr\left(x \in X_i\right)\left(Var\left(y_i\right) + \mathbb{E}\left[\left(\hat{y}_i - \mu_i\right)^2 |m_{it} = r_i\right]\right) \\
&= \sum_{i=1}^k \Pr\left(x \in X_i\right)\left(1 + \frac{1}{r_i}\right) Var\left(y_i\right).
\end{aligned}$$

The first equality uses lemma 8 and the second equality uses the same logic as the proof of lemma 9. Immediately one sees that, for a given categorization, it is optimal to choose to make observations in categories that have a large probability mass and/or a large variance. It is also evident that the marginal benefit of new observations is decreasing.

Forming new categories and concepts is intellectually demanding. More specifically it seems reasonable to assume that this cost is not continuous in the magnitude of the conceptual change; even small changes are likely to involve some strictly positive minimal costs. Thus there is reason not to perform recategorization continuously, rather it is optimal to perform such activities occasionally. Directly after a recategorization the benefit of new observations is high and efforts are rationally directed at performing experiments within the framework of the current categorization. After a while the marginal benefit of new observations has declined sufficiently much to make it more valuable to invest efforts in category formation rather than observation. In this way periods of observation with fixed categories and periods of recategorization will alternate. Note the similarity with how Kuhn (1970) describes the interaction of normal science and scientific revolutions.

4.5. Vagueness. So far it has been assumed that all categorizations $C \in \Psi$ are partitionings of V. Thus all objects in V fall into some category. In the case of categories that are defined by vague concepts this is not a completely innocent assumption. If a concept is vague then there are objects that neither fall under the concept nor fall under its negation. For instance, the concept of a tall person is vague, since there are persons that we have trouble classifying as either tall or not tall. Thus the categories of tall and non-tall person do not jointly exhaust the set of all persons.

Generally, for vague categorizations there are objects that do not belong to any category, but are located in a "no mans land" between categories. For these objects there is no specific category that can serve as basis for predictions. A sensible way to treat such objects is to view them as belonging to a higher level category, corresponding to the whole set V. For instance, a person who is neither tall nor short is still a person. Predictions about such objects are thus based on all objects in the data base. The following proposition establishes that, depending on the category mean, one can decrease or increase expected prediction error by basing the prediction about the category on all objects, using $\hat{y}_t$, rather than basing it just on the objects in that category, using $\hat{y}_{it}$.

PROPOSITION 8. *There exists some finite L' such that if $L > L'$ then for any category $C_i \in C$ with $\mu_i \neq \mu$, expected prediction error is increased by using the prediction $\hat{y}_{it}$ rather than $\hat{y}_t$. And if $L > L'$ then for any category $C_i \in C$ with $\mu_i = \mu$, expected prediction error is decreased by using the prediction $\hat{y}_{it}$ rather than $\hat{y}_t$.*

Thus if there is a category C_i in C whose mean μ_i is equal to the over all mean μ then one can decrease expected prediction error by employing a categorization that is vague in the sense that the objects in C_i do not belong to any category in C. Hence proposition 8 says that it might be optimal to have a partially vague categorization. This can be contrasted with Lipman 2006 who argues that, according to standard models, vagueness is *never* optimal.

5. Related Literature

5.1. Formal Theories of Categorization . Fryer and Jackson (2008) consider a notion of optimal categorization. The model of their important paper has many similarities with the present model; objects are represented as vectors in some space of features, and the prediction about a new object in a category is based on the average of past objects in that category. But there are also some important differences: First, the number of categories is exogenously given. Second, although the purpose of categorization is to generate predictions Fryer and Jackson do not define optimality in terms of minimization of prediction error. Instead they define the optimal categorization as the one that minimizes the sum of within-category differences between objects that have already been encountered. Third, the probability of encountering different objects is not modeled.[16] As a consequence the trade-off that is central to the present paper, cannot be formulated within their framework. Also they can not explore the comparative statics that I do.

After the first version of the present paper was written (and presented in Copenhagen, November 14-15, 2008) it came to my attention that Al-Najjar and Pai (first version December 2008) have developed a model of coarse decision making, which discusses categorization. Their set up is somewhat different in some technical aspects, for instance they consider a finite space of objects, and a finite set of categorizations. They define optimality in a different way than here (using the supremum norm). Their model builds on so-called Vapnik Chervonenkis theory while the present paper only uses basic statistics and probability theory. Regarding results, Al-Najjar and Pai show that when data is scarce then the optimal categorization has less categories than objects (data is always scarce in my model since V is infinite and the data base finite). However, they do not provide any results on what the categorization should look like, and they do not perform any comparative static analysis.

Another related paper is Peski (2007). He intends to explain why categorization may be an optimal way to make predictions – i.e. the question is not what the optimal categorization looks like. But like in the present paper there is a trade-off between fitting and over-fitting. However Peski's model makes some very restrictive assumptions: First of all, Peski argues for the usefulness of categorization by comparing a subject who makes predictions based on categorization with a subject who uses Bayesian updating. The Bayesian subject's prior over the states of the world is symmetric, in

[16] When relating their optimality criterion to utility maximization, they simply assume that the distribution of future objects will be the same as the empirical distribution of already encountered objects.

the sense that the prior is invariant with respect to relabeling of objects and properties. Under this assumption the predictions of the categorizing subject asymptotically approaches the predictions of the Bayesian subject. However, in order for it to be sensible to define optimality in terms of what the Bayesian predicts, the symmetric prior must be objectively correct. This is an extreme assumption. (Also note that this result only holds asymptotically whereas I define optimality for any finite number of observations.) Second, properties are modeled as being discrete. Consequently object similarity is measured as the number of shared properties, and there is no notion of similarity between objects that do not share a property. In reality many properties come in degrees and people are able to judge similarity of properties. For instance a yellow object is judged to be more similar to an orange object than a blue object, and this corresponds to objective similarity in wave length. My purpose is to relate the objective distribution of objects to the way that we slice up the world in categories. It is then crucial to acknowledge that objects and properties are not uniformly distributed. It is also crucial to allow for similarity of properties. Partly because of this, Peski's model does not allow one to study comparative statics regarding the number of categories.[17]

Mullhainathan (2002) provides a Bayesian model of categorization. There is an exogenously given set of types of objects. Each type is associated with a probability distribution over outcomes. A subject chooses a proper subset of these distributions, which correspond to the set of categories used. When a new object is encountered the subject pick the category that is most likely given the data she already has about that object. Then the distribution associated with that category is used for predictions about the object in question. There are many differences compared to the present paper: There is an exogenous set of "true" categories and the subject has a prior about how objects are distributed within categories. Related to this, objects are allocated to categories on the basis of Bayesian inference.

Jehiel (2005) develops a notion of analogy based expectations equilibrium for extensive form games. Players bundle together the nodes of the opponents into analogy classes in order to predict the opponents' behavior. A player expects the same behavior in all nodes in an analogy class. In equilibrium these expectations are correct on average. The equilibrium is parameterized by the analogy classes, which are exogenous. Jehiel and Samet (2007) define a notion of valuation equilibrium. Players bundle their own strategies into different similarity classes, when predicting their own payoffs. The

[17] A more technical difficulty is the sufficient data condition, which states that the number of observations asymptotically becomes infinitely much larger than the number of distinct features in the data base of past observations. Since the number of features is infinite this assumption is quite demanding.

same payoff is expected for each strategy in the same similarity class, and in equilibrium the expectations are correct on average. The similarity classes are exogenous, even though Jehiel and Samet discuss the possibility of endogenizing them.

There is a literature, starting with Dow (1991), that examines the optimal way to partition a state space in the face of limited memory. In these models the number of cells in the partition is determined exogenously by the bound on memory. (See chapter 5 of Rubinstein 1998) for a general discussion of optimal partitions.) The subject is assumed to have a prior defined on the state space. In the categorization model discussed in this paper the subject's prior probabilities are instead generated from observations.

In the psychology literature the closest related theory is the "rational theory" of Anderson (1991). In his model predictions about objects in a given category are based on all objects in all categories, thus making the distinct purpose of categories somewhat difficult to understand. The model is Bayesian and postulates that subjects have a prior about the category structure of the encountered objects. In particular subjects have a prior about how similar object are in order to belong to the same category, as well as a prior about the probability that objects in a certain category exhibit different properties. As a result the categorization depends critically on assumptions about priors. Finally, Anderson unable to formulate the trade off that determines the number of categories according to the present paper.

In the field of machine learning there are several models related to categorization. The approach most relevant to the question of optimality raised in this paper, is cluster analysis (for a review see e.g. Jain et al. 1999). Still, there are some important points of difference. In cluster analysis it is assumed that there is a certain set of distributions generating data, represented as points in some multidimensional space. The test of optimality (if the question is addressed at all) is to be able to distinguish which observations that where drawn from which distribution. The goodness of fit of a model for allocating objects to the different distributions can then be assessed with some information criterion from statistics. In this sense the approach assumes the existence of a given number of natural kinds, whereas mine does not. Another important difference is that optimality is only evaluated with respect to the variables that are used to define clusters. This means that optimality is not evaluated with respect a variable that is not observed when categorizing a new object.

5.2. Similarity-Based Predictions . Gilboa and Schmeidler (1995) develop a model of case-based decision making, which Gilboa and Schmeidler (2003) adapt to prediction problems; given a data base of past cases the subject's task is to rank the likelihood of different outcomes in a new case. Gilboa et al. (2006) provide an axiomatization of a similarity based prediction rule for the case of predicting a real-valued variable y. The prediction rule presumes the existence of a measure of similarity between cases. It states that the value of y in the case at hand will be equal to the similarity weighted average of that variable in the past cases.[18] The axiomatization assures that there exists such a similarity function if and only if the probability rankings made by the subject, given various data bases, satisfies certain axioms. However, the axiomatization only tells us that a similarity function exists, not what it looks like.

This approach, in particular Gilboa et al. (2006), is related to the present paper in an interesting way. One way of phrasing the difference is to say that I consider a certain subset of similarity functions, namely category-based similarity functions. This is the set of functions that treat all cases in the same category as exactly similar to each other and treat a case in a category as completely dissimilar to any case outside that category. Furthermore, instead of axiomatizing this category-based similarity function I derive the optimal such similarity function. In discussing their similarity based approach to generation of priors, Gilboa et al. (2008) say: "An obvious question about this approach is that it may appear that the problem of finding an appropriate probability has simply been replaced by the problem of finding an appropriate similarity function" (p. 185). They suggest that this can be done empirically. The present paper instead tries to narrow down the problem by first acknowledging the importance of categorical reasoning, and then looking for the optimal way of forming categories.[19]

6. Conclusion

I have provided a framework for the study of optimal categorization for the purpose of making predictions. The optimal number of categories is endogenous to the model. A small category results in smaller variance of objects in that category. A large category leads to a large number of experienced objects in the category, thus improving the precision of the predictions of the category mean. Thus the optimal categorization strikes a balance between fitting and over-fitting. This can explain the fact that the privileged level of categorization – the so-called basic level – is neither the coarsest nor the finest one. Comparative statics yield several predictions about how the optimal categorization varies with the number of observations and the distribution of objects.

[18] Similarly Billot et al (2005) axiomatize a similarity based prediction rule for the case of predicting a multi-dimensional probability vector. The rule states that the probability of obtaining a particular outcome in a new case is equal to the similarity weighted frequency of that outcome in the past.

[19] Gilboa et al. (forthcoming) extend the methodology to the problem of making a prediction in the form of a density function over a set of real values. This is done with a Kernel estimation approach where the Kernel depends on the values of the cases in the data base, in a way that is similar to the similarity weighting in the point prediction case. It would be interesting to extend the model of the present paper to the case of predictions in the form of densities.

It would be interesting to test experimentally some of the predictions of the model that have not been tested before, such as the predictions that the optimal number of categories are increasing in the variance of the density. The model is simple but the insights should carry over to more general settings, e.g. multidimensional objects and predictions of discrete variables. Also the model could be applied to categorization of games and strategies – for instance, one could define solution concepts for agents who form beliefs based on (optimal) categorization.

Appendix: Proofs

Proof of Lemma 8. We have

$$\begin{aligned} EPE\left(C, v^{t-1}\right) &= \sum_{i=1}^{k} \int_{(x,y) \quad \in C_i} f\left(x,y\right)\left(y-\hat{y}_{it}\right)^2 d\left(x,y\right) \\ &= \sum_{i=1}^{k} \int_{y\in Y} \int_{x \in X_i} f\left(x,y\right) dx \left(y-\hat{y}_{it}\right)^2 dy \\ &= \sum_{i=1}^{k} \int_{y\in Y} \Pr\left(x\in X_i\right) f\left(y|x\in X_i\right)\left(y-\hat{y}_{it}\right)^2 dy, \end{aligned}$$

where the last equality uses the definition of $f\left(y|x\in X_i\right)$. Note that

$$\left(y-\hat{y}_{it}\right)^2 = \left(y-\mu_i\right)^2 + \left(\hat{y}_{it}-\mu_i\right)^2 - 2\left(y-\mu_i\right)\left(\hat{y}_{it}-\mu_i\right).$$

Using this we have

$$\begin{aligned} EPE\left(C, v^{t-1}\right) = &\sum_{i=1}^{k} \Pr\left(x\in X_i\right)\left(\int_{y\in Y} f\left(y|x\in X_i\right)\left(y-\mu_i\right)^2 dy + \left(\hat{y}_{it}-\mu_i\right)^2\right) \\ &- \sum_{i=1}^{k} \Pr\left(x\in X_i\right) 2 \left(\int_{y\in Y} f\left(y|x\in X_i\right) y dy - \mu_i\right)\left(\hat{y}_{it}-\mu_i\right) \end{aligned}$$

The desired result follows from the facts that the second factor on the right hand side is equal to zero, and

$$\int_{y\in Y} f\left(y|x\in X_i\right)\left(y-\mu_i\right)^2 dy = Var\left(y|x\in X_i\right) = Var\left(y_i\right).$$

$\square$

Proof of Lemma 9. We have

$$\begin{aligned} EPE\left(C,t\right) &= \mathbb{E}\left[EPE\left(C, v^{t-1}\right)\right] \\ &= \sum_{i=1}^{k} \Pr\left(x\in X_i\right) Var\left(y_i\right) \\ &+ \sum_{i=1}^{k} \Pr\left(x\in X_i\right) \sum_{r=1}^{t-1} \Pr\left(m_{it}=r\right) \mathbb{E}\left[\left(\hat{y}_{it}-\mu_i\right)^2 | m_{it}=r\right] \\ &+ \sum_{i=1}^{k} \Pr\left(x\in X_i\right) \Pr\left(m_{it}=0\right) \mathbb{E}\left[\left(\hat{y}_{t}-\mu_i\right)^2 | m_{it}=0\right]. \end{aligned}$$

The number of objects in a category, m_{it}, has a binomial distribution as follows

$$\Pr(m_{it} = r) = \frac{(t-1)!}{r!\,(t-1-r)!} \left(\Pr(x \in X_i)\right)^r \left(1 - \Pr(x \in X_i)\right)^{t-1-r}.$$

If $r > 0$ then $E[\hat{y}_{it} | m_{it} = r] = \mu_i$, so

$$\begin{aligned}
\mathbb{E}\left[(\hat{y}_{it} - \mu_i)^2 | m_{it} = r\right] &= Var(\hat{y}_{it} | m_{it} = r) \\
&= \sum_{j=1}^{r} \frac{1}{r^2} Var(y_i | m_{it} = r) \\
&= \sum_{j=1}^{r} \frac{1}{r^2} Var(y_i) \\
&= \frac{1}{r} Var(y_i).
\end{aligned}$$

Plugging this into the expression above yields the desired result. □

Proof of Lemma 10. The equality $\mathbb{E}[Var(y|x)] = Var(\mathbb{E}[y|x])$ is a standard result. Conditioning on $x \in X_i$ is straightforward. A full proof can be obtained from the author upon request. □

Proof of Lemma 11. Assume $k < \kappa < \infty$ and assume that each X_i is the union of at most ι intervals. Any categorization C with k categories can be described by a set of $\kappa\iota - 1$ points on $[a, b]$ together with a mapping from the induced $(\kappa\iota)$ subintervals to the set $\{1, 2, ..., k\}$. Take any mapping $\nu$from subintervals to $\{1, 2, ..., k\}$. Choosing an optimal categorization among the categorizations that are consistent with the mapping $\nu$is equivalent to choosing a point z in the compact set

$$Z = \left\{z \in [a, b]^{\kappa\iota - 1} : z_j \leq z_{j+1} \forall j \in \{1, ..., \kappa\iota - 2\}\right\}$$

in order to minimize the objective function $EPE(C, v^{t-1})$. Furthermore, since f is continuous in x, the objective function is continuous in z. Hence by Weierstrass' maximum theorem there exists a solution $z^*(\nu)$. This was for a given mapping ν from subintervals to $\{1, 2, ..., k\}$. Since there are only a finite number of mappings from $\kappa\iota$ subintervals to the set $\{1, 2, ..., k\}$. The desired result follows. □

The following lemma is needed for the proof of proposition 1.

Lemma 12. *Consider categorizations with $k/t \geq \gamma > 0$. There is some $\xi > 0$ such that*

$$\sum_{i=1}^{k} \Pr(x \in X_i) \Pr(m_{it} = 0) > \xi.$$

Proof of Lemma 12. Let $p_{\max} = \max_i \Pr(x \in X_i)$ and $p_{\min} = \min_i \Pr(x \in X_i)$. Note that $p_{\max} < 1 - (k-1)\, p_{\min}$ and since $p_{\min} > \rho p_{\max}$, we have $p_{\max} < 1 - (k-1)\, \rho p_{\max}$ or equivalently

$$p_{\max} < \frac{1}{((k-1)\,\rho + 1)} = \frac{1}{k\rho - \rho + 1}.$$

Since $\rho \in (0,1)$ this implies $p_{\max} < 1/k\rho$. Thus

$$\begin{aligned}
\sum_{i=1}^{k} \Pr(x \in X_i) \Pr(m_{it} = 0) &\geq (1 - p_{\max})^{t-1} \\
&> \left(1 - \frac{1}{k\rho}\right)^{t-1} \\
&\geq \left(1 - \frac{1}{k\rho}\right)^{\frac{k}{\gamma}-1} \\
&= \left(\left(1 - \frac{1}{k\rho}\right)^{k\rho}\right)^{\frac{1}{\gamma\rho}} \left(1 - \frac{1}{k\rho}\right)^{-1}.
\end{aligned}$$

Where the last inequality uses $t \leq k/\gamma$. We have

$$\lim_{k\to\infty} \left(\left(1 - \frac{1}{k\rho}\right)^{k\rho}\right)^{\frac{1}{\gamma\rho}} \left(1 - \frac{1}{k\rho}\right)^{-1} = \left(\lim_{k\to\infty} \left(1 - \frac{1}{k\rho}\right)^{k\rho}\right)^{\frac{1}{\gamma\rho}} = e^{-\frac{1}{\gamma\rho}},$$

so

$$\sum_{i=1}^{k} \Pr(x \in X_i) \sum_{r=1}^{t-1} \Pr(m_{it} = r) \frac{1}{r} > e^{-\frac{1}{\gamma\rho}},$$

for all t and all categorizations satisfying $k = t$. For all $\rho > 0$, $e^{-\frac{1}{\gamma\rho}}$ is strictly positive.[20]

□

Proof of Proposition 1. **(a)** **(i)** Write

$$\sum_{i=1}^{k} \Pr(x \in X_i)\, Var(y_i) = \sum_{i=1}^{k} \left(\int_{x\in X_i} f(x)\, dx\right) \int_{y\in Y} f(y|x \in X_i)(y - \mu_i)^2\, dy.$$

For any t, let all sets X_i be intervals of length $(b-a)/k$. If $k \to \infty$ then the above expression approaches

$$\int_{x\in X} f(x) \left(\int_{y\in Y} f(y|x)(y - \mathbb{E}(y|x))^2\, dy\right) dx = \int_{x\in X} f(x)\, Var(y|x)\, dx.$$

[20] The lemma would not hold if we did not impose the restriction that there is some $\rho(0,1)$ such that $\Pr(x \in X_i) / \Pr(x \in X_i) > \rho$ for all i and j. Roughly speaking the reason is that one could then let some categories go to zero much faster than other categories so that the expected number of objects in the slowly decreasing categories goes to infinity.

Hence for any $\varepsilon > 0$ there is a finite k' such that for any $k > k'$ (and any t) one can partition X such that

$$\left|\sum_{i=1}^{k} \Pr(x \in X_i) Var(y_i) - \int_{x \in X} f(x) Var(y|x) dx\right| < \varepsilon.$$

Thus for any $\varepsilon > 0$ there is a finite k' such that for any $k > k'$ (and any L and T) one can partition X such that

$$\left|\frac{1}{T-L} \sum_{t=L+1}^{T} \sum_{i=1}^{k} \Pr(x \in X_i) Var(y_i) - \int_{x \in X} f(x) Var(y|x) dx\right| < \varepsilon.$$

(ii) Consider $EPE(C,t)$ and fix k. If one lets $t \to \infty$ so that $k/t \to 0$, then one obtains

$$\sum_{r=1}^{t-1} \Pr(m_{it} = r) \frac{1}{r} \to 0,$$

for all categories, and hence

$$EPE(C,t) \to \sum_{i=1}^{k} \Pr(x \in X_i) Var(y_i),$$

so for any $\varepsilon > 0$ and any finite k, there is a finite t', such that if $t > t'$ then one can partition X in a way such that

$$\left|EPE(C,t) - \sum_{i=1}^{k} \Pr(x \in X_i) Var(y_i)\right| < \varepsilon.$$

Thus for any $\varepsilon > 0$ and any finite k there are finite L' and T' such that if $L > L'$ or $T > T'$ then

$$\left|EPE(C,T,L) - \frac{1}{T-L} \sum_{t=L+1}^{T} \sum_{i=1}^{k} \Pr(x \in X_i) Var(y_i)\right| < \varepsilon.$$

(iii) From (i) and (ii) it follows that for any $\varepsilon > 0$ there are finite numbers k', L' and T' such that if $k > k'$ and if $L > L'$ or $T > T'$, then one can partition X in a way such that

$$\left|EPE(C,T,L) - \int_{x \in X} f(x) Var(y|x) dx\right| < 2\varepsilon.$$

(iv) Now I show that one cannot obtain such a low expected prediction error unless $k < L$. Consider $EPE(C,t)$ and restrict attention to the set of categorizations with $k \geq t$. We have $\min_{x \in X} Var(y|x) > \mathbb{E}\left[(\hat{y}_t - \mu_i)^2 | m_{it} = 0\right]$, so

$$EPE(C,t) > \sum_{i=1}^{k} \Pr(x \in X_i) Var(y_i) + \left(\min_{x \in X} Var(y|x)\right) \sum_{i=1}^{k} \Pr(x \in X_i) \Pr(m_{it} = 0).$$

Let $k \to \infty$, implying $t \to \infty$, and let $\Pr(x \in X_i) \to 0$ for all categories. Then, by (i) and lemma 12 there is some $\xi > 0$ such that

$$EPE(C,t) > \int_{x\in X} f(x)\, Var(y|x)\, dx + \xi \min_{x\in X} Var(y|x).$$

Hence, there is some ε such that for any L, and any categorization with $k \geq L$ categories, we have

$$EPE(C,T,L) - \int_{x\in X} f(x)\, Var(y|x)\, dx > 2\varepsilon.$$

(b) The proof of part (b) is very similar to the proof of part (a), and therefore omitted.

(c) First consider minimization of $EPE(C,t)$ rather than $EPE(C,T,L)$. Set $k = \sqrt{t}$. If $t \to \infty$ then $k = \sqrt{t} \to \infty$ and $k/t = 1/\sqrt{t} \to 0$. Thus if $t \to \infty$ then one can let $k \to \infty$ so that, according to (i) above,

$$\sum_{i=1}^{k} \Pr(x \in X_i)\, Var(y_i) \to \int_{x\in X} f(x)\, Var(y|x)\, dx,$$

and still have $k/t \to \infty$, so that according to (ii)

$$EPE(C,t) \to \sum_{i=1}^{k} \Pr(x \in X_i)\, Var(y_i),$$

implying

$$EPE(C,t) \to \int_{x\in X} f(x)\, Var(y|x)\, dx.$$

If instead one sets a strictly positive lower bound on k/t then according to lemma 12 there is some $\varepsilon > 0$ such that

$$\min_{C\in\Psi} EPE(C,t) \to \int_{x\in X} f(x)\, Var(y|x)\, dx + \varepsilon,$$

as $t \to \infty$. Thus as $t \to \infty$, $EPE(C,t)$ is minimized by letting $k/t \to 0$. Hence for any L, as $T \to \infty$, $EPE(C,T,L)$ is minimized by letting $k/T \to 0$. □

Proof of Corollary 1. According to proposition 1 the number of categories is optimally lower than T. Since T is finite this number can be used instead of κ in the proof of lemma 11. □

The proof of proposition 2 relies on the following lemma.

Lemma 13. *Let A and B be disjoint intervals with $\Pr(x \in A) > 0$, and $\Pr(x \in B) > 0$. We have $E(y|x \in A) \neq E(y|x \in B)$ if and only if*

$$\Pr(x \in A \cup B)\, Var(y|x \in A \cup B) - \sum_{I\in\{A,B\}} \Pr(x \in I)\, Var(y|x \in I) > 0.$$

Proof of Lemma 13. Recall
$$f\left(y|x\in A\cup B\right)=\frac{\Pr\left(y\text{ and }x\in A\cup B\right)}{\Pr\left(x\in A\cup B\right)},$$
so
$$\begin{aligned}&\Pr\left(x\in A\cup B\right)Var\left(y|x\in A\cup B\right)\\&=\int_{y\in Y}\Pr\left(x\in A\cup B\right)f\left(y|x\in A\cup B\right)\left(y-\mathbb{E}\left(y|x\in A\cup B\right)\right)^2dy\\&=\int_{y\in Y}\Pr\left(y\text{ and }x\in A\cup B\right)\left(y-\mathbb{E}\left(y|x\in A\cup B\right)\right)^2dy.\end{aligned}$$
Similarly for $I\in\{A,B\}$
$$\Pr\left(x\in I\right)Var\left(y|x\in I\right)=\int_{y\in Y}\Pr\left(y\text{ and }x\in I\right)\left(y-\mathbb{E}\left(y|x\in I\right)\right)^2dy.$$
Putting this together, and using
$$\Pr\left(y\text{ and }x\in I\right)=\int_{x\in I}f\left(x,y\right)dx,$$
yields
$$\begin{aligned}&\Pr\left(x\in A\cup B\right)Var\left(y|x\in A\cup B\right)-\sum_{I\in\{A,B\}}\Pr\left(x\in I\right)Var\left(y|x\in I\right)\\&=\int_{y\in Y}\Pr\left(y\text{ and }x\in A\cup B\right)\left(y-\mathbb{E}\left(y|x\in A\cup B\right)\right)^2dy\\&-\sum_{I\in\{A,B\}}\int_{y\in Y}\Pr\left(y\text{ and }x\in I\right)\left(y-\mathbb{E}\left(y|x\in I\right)\right)^2dy\\&=\int_{y\in Y}\int_{x\in A\cup B}f\left(x,y\right)dx\left(y-\mathbb{E}\left(y|x\in A\cup B\right)\right)^2dy\\&-\sum_{I\in\{A,B\}}\int_{y\in Y}\int_{x\in I}f\left(x,y\right)dx\left(y-\mathbb{E}\left(y|x\in I\right)\right)^2dy\\&=\sum_{I\in\{A,B\}}\int_{y\in Y}\int_{x\in I}f\left(x,y\right)dx\left(\left(y-\mathbb{E}\left(y|x\in A\cup B\right)\right)^2dy-\left(y-\mathbb{E}\left(y|x\in I\right)\right)^2\right)dy\\&\geq 0,\end{aligned}$$
where the weak inequality follows from the fact that the function
$$q\left(z\right)=\int_{y\in Y}\int_{x\in T}f\left(x,y\right)dx\left(y-z\right)^2dy,$$
is minimized at $z=E\left(y|x\in I\right)$. The inequality is strict if $E\left(y|x\in A\right)\neq E\left(y|x\in B\right)$.

$\square$

PROOF OF PROPOSITION 2. Compare categorizations $C' = \{C_\alpha, C_\beta\}$ and $C'' = \{C_\gamma\}$ such that in C'' all objects belong to the same category, $C_\gamma = V$, while in C' objects from E belong to a category of its own, $C_\alpha = E \times Y$, and all other objects belong to a category $C_\beta = F \times Y$. We know from before (proof of proposition 1) that for any $\varepsilon > 0$ and any finite k there is a finite t' such that if $t > t'$ then

$$EPE\left(C', t\right) - \left(\Pr\left(x \in E\right) Var\left(y_\alpha\right) + \Pr\left(x \in F\right) Var\left(y_\beta\right)\right) < \varepsilon,$$

and

$$EPE\left(C'', t\right) - \Pr\left(x \in E \cup F\right) Var\left(y_\gamma\right) < \varepsilon,$$

so if $t > t'$, then

$$\begin{aligned} EPE\left(C', t\right) - EPE\left(C'', t\right) < {} & \Pr\left(x \in E\right) Var\left(y_\alpha\right) + \Pr\left(x \in F\right) Var\left(y_\beta\right) \\ & - \Pr\left(x \in E \cup F\right) Var\left(y_\gamma\right) + 2\varepsilon. \end{aligned}$$

It follows from lemma 13 that if $E\left(y|x \in E\right) \neq E\left(y|x \in F\right)$ then

$$\Pr\left(x \in E\right) Var\left(y_\alpha\right) + \Pr\left(x \in F\right) Var\left(y_\beta\right) - \Pr\left(x \in E \cup F\right) Var\left(y_\gamma\right) < 0.$$

Hence, for sufficiently large t' (and hence sufficiently small ε) we have $EPE\left(C', t\right) < EPE\left(C'', t\right)$. It follows that it L is sufficiently large or if T is sufficiently large, then $EPE\left(C', T, L\right) < EPE\left(C'', T, L\right)$. □

PROOF OF PROPOSITION 3. There are $k \leq t - 1$ categories. If $f\left(y|x\right) = f\left(y|x'\right)$ for all $x, x' \in X$, then $Var\left(y_i\right) = Var\left(y\right)$, and

$$\int_{y \in Y} f\left(y|x \in X_i\right)\left(y - \mu\right)^2 dy = \int_{y \in Y} f\left(y\right)\left(y - \mu\right)^2 dy = Var\left(y\right).$$

Furthermore $\mathbb{E}\left(\left(\hat{y}_t - \mu\right)^2 | m_{it} = 0\right) = 0$, so

$$EPE\left(C, t\right) = Var\left(y\right)\left(1 + \sum_{i=1}^{k} \Pr\left(x \in X_i\right) \sum_{r=1}^{t-1} \Pr\left(m_{it} = r\right) \frac{1}{r}\right).$$

Suppose categorization C' has $k - 1$ concepts and that categorization C'' is identical except that category $k - 1$ in C' is divided into two categories so that categorization C'' has k concepts. Let X_i' and X_i'' denote the projection to the x-axis, of category C_i in categorization C' and C'', respectively. Similarly let m_{it}' and m_{it}'' denote the number of objects in category i at date t, for categorization C' and C'', respectively. We have

$$\begin{aligned}
&EPE\left(C'',t\right)-EPE\left(C',t\right)\\
&=Var\left(y\right)\Pr\left(x\in X''_k\right)\sum_{s=1}^{t-1}\Pr\left(m''_{kt}=s\right)\frac{1}{s}\\
&+Var\left(y\right)\Pr\left(x\in X''_{k-1}\right)\sum_{s=1}^{t-1}\Pr\left(m''_{(k-1)t}=s\right)\frac{1}{s}\\
&-Var\left(y\right)\Pr\left(x\in X'_{k-1}\right)\sum_{s=1}^{t-1}\Pr\left(m'_{(k-1)t}=s\right)\frac{1}{s}\\
&=Var\left(y\right)\Pr\left(x\in X''_k\right)\left(\sum_{s=1}^{t-1}\Pr\left(m''_{kt}=s\right)\frac{1}{s}-\sum_{s=1}^{t-1}\Pr\left(m'_{(k-1)t}=s\right)\frac{1}{s}\right)\\
&+Var\left(y\right)\Pr\left(x\in X''_{k-1}\right)\left(\sum_{s=1}^{t-1}\Pr\left(m''_{(k-1)t}=s\right)\frac{1}{s}-\sum_{s=1}^{t-1}\Pr\left(m'_{(k-1)t}=s\right)\frac{1}{s}\right)\\
&<0.
\end{aligned}$$

□

Proof of Proposition 4. The proof is brief since it uses a logic similar to that in the proof of proposition 1. Let C' be the categorization, with k_1 categories, that minimizes $EPE\left(C,t_1\right)$. To see what categorization C that minimizes $EPE\left(C,t_2\right)$, first suppose, as a benchmark, that one uses the categorization C', and obtains $EPE\left(C',t_2\right)$. By choosing a categorization C'' with a larger number of categories one can reduce

$$\sum_{i=1}^{k}\Pr\left(x\in X_i\right)Var\left(y_i\right).$$

If t_2-t_1 is large enough then one can increase k and still decrease both

$$\sum_{r=1}^{t_2-1}\Pr\left(m_{it_2}=r\right)\frac{1}{r},$$

and

$$\sum_{i=1}^{k}\Pr\left(x\in X_i\right)\Pr\left(m_{it_2}=0\right)\mathbb{E}\left[\left(\hat{y}_{t_2}-\mu_i\right)^2|m_{it_2}=0\right].$$

Thus if t_2-t_1 is large enough then one can increase k and thereby obtain

$$EPE\left(C'',t_1\right)<EPE\left(C',t_1\right).$$

Conversely if $t_2 - t_1$ is large enough then one cannot gain by reducing the number of categories and set $k_2 < k_1$. It is straightforward to translate this reasoning about $t_2 - t_1$ into statements about $L_2 - L_1$ and $T_2 - T_1$. $\square$

Proof of Proposition 5. Restrict attention to the expected prediction error in the set E, denoted $EPE_E(C,t)$. Write $EPE_{E,f}(C,t)$ to make the dependence upon f explicit. Suppose C' is an optimal categorization of E at date t given f_0, i.e $C' \in \arg\min_{C\in\Psi} EPE_{E,f_0}(C,t)$, and suppose that there is no other optimal categorization with a lower number of categories. This categorization C' strikes an optimal balance between the goal of having a few large categories in order to minimize the factors $\sum_{r=1}^{t-1} \Pr(m_{it} = r)\frac{1}{r}$ and $\Pr(m_{it} = 0)$ (one of each for each category), and the goal of having many small categories in order to minimize the factors $Var(y_i)$ (one for each category). Decreasing the number of categories will lead to an increase in at least some of the factors $Var(y_i)$ and a decrease in at least some of the factors $\sum_{r=1}^{t-1} \Pr(m_{it} = r)\frac{1}{r}$ and $\Pr(m_{it} = 0)$. The former effect will dominate the latter so that the total effect will be an increase in prediction error – otherwise C' would not be an optimal categorization with a minimal number of categories. (The effect on the factors $E\left[(\hat{y}_t - \mu_i)^2 | m_{it} = 0\right]$ of increasing the number of categories is ambiguous, but if these terms are decreased by increasing the number of categories it still must be the case that the total effect on expected prediction error, of increasing the number of categories, is positive.)

Now suppose one uses the same categorization C' when the distribution is f_1 (rather than f_0). Then all the factors $\sum_{r=1}^{t-1} \Pr(m_{it} = r)\frac{1}{r}$ and $\Pr(m_{it} = 0)$ are smaller under f_1 than under f_0. But we have

$$\frac{f(x)}{\Pr(x \in X_i)} = \frac{f(x)}{\int_{x\in X_i} f(x)\,dx},$$

and

$$\frac{f_1(x)}{\int_{x\in X_i} f_1(x)\,dx} = \frac{\alpha f_0(x)}{\int_{x\in X_i} \alpha f_0(x)\,dx} = \frac{f_0(x)}{\int_{x\in X_i} f_0(x)\,dx}.$$

so from the expressions for $E[Var(y|x)\,|x \in X_i]$ and $Var(\mathbb{E}[y|x]\,|x \in X_i)$ (and lemma 10) one sees that all the factors $Var(y_i)$ and $E\left[(\hat{y}_t - \mu_i)^2 | m_{it} = 0\right]$ are the same under f_0 and f_1. Also all the factors $E\left[(\hat{y}_t - \mu_i)^2 | m_{it} = 0\right]$ are unaffected. Hence, keeping C' fixed, the only difference between f_0 and f_1 is that the factors $\sum_{r=1}^{t-1} \Pr(m_{it} = r)\frac{1}{r}$ are smaller under f_1 than under f_0. Since it was suboptimal to decrease the number of categories relative to C' under f_0 it must be (even more) suboptimal to decrease the number of categories relative to C' under f_1. $\square$

Proof of Proposition 6. If L and T are sufficiently large then we can neglect the probability of empty categories. Write $EPE_f(C,t)$ to make the dependence on f explicit. Suppose C' is an optimal categorization at date t given f_0, i.e $C' \in \arg\min_{C\in\Psi} EPE_{f_0}(C,t)$, and suppose that there is no other optimal categorization with a lower number of categories. This categorization C' strikes an optimal balance between the goal of having a few large categories in order to minimize the factors $\sum_{r=1}^{t-1}\Pr(m_{it}=r)\frac{1}{r}$ and the goal of having many small categories in order to minimize the factors $Var(y_i)$. Decreasing the number of categories will lead to an increase in at least some of the factors $Var(y_i)$ and a decrease in at least some of the factors $\sum_{r=1}^{t-1}\Pr(m_{it}=r)\frac{1}{r}$. The former will dominate the latter so that the total effect will be an increase in prediction error – otherwise C' would not be an optimal categorization with a minimal number of categories.

If one uses the same categorization C' for f_1 then all the factors $\sum_{r=1}^{t-1}\Pr(m_{it}=r)\frac{1}{r}$ are the same under f_1 as under f_0. Also all the factors $Var(\mathbb{E}[y|x]|x\in X_i)$ are the same under f_1 as under f_0. But we have

$$\mathbb{E}_{f_1}[Var(y|x)|x\in X_i] > \int_{x\in X_i}\frac{f(x)}{\Pr(x\in X_i)}Var_{f_0}(y|x)\,dx = \mathbb{E}_{f_0}[Var(y|x)|x\in X_i],$$

for all categories in C', so $Var_{f_0}(y_i) < Var_{f_1}(y_i)$ for all categories in C'. Hence, when C' is kept the same then the only difference between f_0 and f_1 is that the factors $Var(y_i)$ are larger under f_1 than under f_0. Disregarding the probability of empty categories we have

$$\begin{aligned}EPE_{f_1}(C,t) &= EPE_{f_0}(C,t)\\ &\quad+\sum_{i=1}^{k}\Pr(x\in X_i)\left(\mathbb{E}_{f_1}[Var(y|x)|x\in X_i]-\mathbb{E}_{f_0}[Var(y|x)|x\in X_i]\right)\\ &\quad\times\left(1+\sum_{r=1}^{t-1}\Pr(m_{it}=r)\frac{1}{r}\right).\end{aligned}$$

By assumption $EPE_{f_0}(C,t)$ is minimized by C'. The second term is strictly increasing in the number of categories. In total the minimal optimal number of categories will be weakly lower under f_1 than under f_0. □

Proof of Proposition 7. A longer version of this proof can be obtained from the author upon request. Note that the assumption that x is uniformly distributed on X, implies that $f(x)=1$ for all $x\in X$, and hence that $f(x,y)=f(y|x)$. Then derive the variance of y in interval $A_i=[a_i,b_i)$. We have

$$\Pr(x\in A_i)=b_i-a_i,$$

and

$$f(y|x \in A_i) = \frac{\int_{x\in A} f(y,x)\,dx}{\Pr(x \in A_i)} = \frac{1}{b_i - a_i}\int_{x\in A_i} f(y|x)\,dx,$$

and

$$\mathbb{E}(y|x \in A_i) = \left(\alpha + \beta\frac{(a_i + b_i)}{2}\right).$$

Using this we get, after a fair amount of manipulation,

$$Var(y_i) = \frac{\beta^2 (b_i - a_i)^2}{12} + \sigma^2. \tag{6.1}$$

(a) Now we show that the optimal categories are intervals on the x-axis. Take a categorization C where not all categories are convex. That means that without loss of generality one can assume that there is a category C_α such that $X_\alpha = \cup_{s=1}^{S}[a_s, b_s)$, with $b_s < a_{s+1}$. Let $EPE_\alpha(C,t)$ denote the expected prediction error for objects in this category;

$$EPE_\alpha(C,t) = Var(y_\alpha)\left(1 + \sum_{r=0}^{t-1}\Pr(m_{\alpha t} = r)\frac{1}{r+1}\right).$$

Consider a categorization C' that is a modification of C such that $X'_\beta = [a_1, b)$ where $b = a_1 + \sum_{s=1}^{S}(b_s - a_s)$. The other categories are only moved to the right so that if, under categorization C the point $p > a_1$ was a boundary point between two categories then, under categorization C' this boundary is located at the point $p + \sum_{s=1}^{S}(b_s - a_s)$. Let $EPE_\beta(C',t)$ denote the expected prediction error for objects in category $C_\beta \in C'$;

$$EPE_\beta(C',t) = Var(y_\beta)\left(1 + \sum_{r=0}^{t-1}\Pr(m_{\beta t} = r)\frac{1}{r+1}\right).$$

Comparing these expressions one finds $EPE_\beta(C',t) < EPE_\alpha(C,t)$. From equation 6.1 we see that the expected prediction error for objects in the other categories are unaffected so $EPE(C',t) < EPE(C,t)$. Hence the categorization with a convex category is better than the one with a non-convex category.

We know that an optimal categorization with k concepts has $X_i = [a_i, b_i)$ for $i \in \{1,...,k-1\}$ and $X_k = [a_k, b_k] = [a_k, 1]$. Letting $d_i = b_i - a_i$, we seek a categorization that minimizes

$$EPE(C,t) = \sum_{i=1}^{k}(b_i - a_i)\left(\frac{\beta^2(b_i - a_i)^2}{12} + \sigma^2\right)\left(1 + \sum_{r=0}^{t-1}\Pr(m_{it} = r)\frac{1}{r+1}\right),$$

where

$$\Pr(m_{it} = r) = \frac{(t-1)!}{r!\,(t-1-r)!}\left((b_i - a_i)\right)^r\left(1 - (b_i - a_i)\right)^{t-1-r}.$$

Since $EPE\left(C,t\right)$ is quadratic in $b_i - a_i$ it is optimal to have $b_i - a_i = 1/k$ for all i. Since we have assumed $T - L = 1$ this finishes the proof of (a).

(b) Let $b_i - a_i = d$ for all i. The task is then to minimize

$$EPE\left(C,t\right) = \left(\frac{\beta^2 d^2}{12} + \sigma^2\right)\left(1 + \sum_{r=1}^{t-1} \frac{(t-1)!}{r!\left(t-1-r\right)!}\left(d\right)^r\left(1-d\right)^{t-1-r}\frac{1}{r}\right),$$

where I have used the fact that $k = 1/d$. The first order condition (FOC) of $EPE\left(C,t\right)$ w.r.t. d is

$$\left(\frac{\beta^2}{6}\right)\left(d + \sum_{r=1}^{t-1}\left(d - \left(\frac{1}{2} + \sigma^2\right)\frac{d\left(d\left(t-1\right)-r\right)}{\left(1-d\right)}\right)\frac{1}{r}\Pr\left(m_{it} = r\right)\right) = 0.$$

It can be verified that there is some finite t' such if $t > t'$ then the second order condition (SOC) for an interior minimum is satisfied. From the expression for the second order condition one also finds that that for sufficiently large t it is the case that

$$\frac{\partial}{\partial d}\left(d + \sum_{r=1}^{t-1}\left(d - \left(\frac{1}{2} + \sigma^2\right)\frac{d\left(d\left(t-1\right)-r\right)}{\left(1-d\right)}\right)\frac{1}{r}\Pr\left(m_{it} = r\right)\right) > 0,$$

so the left hand side (LHS) of the FOC is increasing in d. Now if we increase β then the LHS increases so in order to satisfy the FOC one has to decrease d. If we increase σ^2 then the LHS decreases and so in order to satisfy the FOC one has to increase d. □

Proof of Proposition 8. **(i)** Suppose that there is a categorization C' such that a subset $A \subseteq X$ is not categorized, in the sense that A has an empty intersection with any category in C''. Then the expected prediction error for objects in A, with expectation taken over the set of objects, is

$$EPE_A\left(C', v^{t-1}\right) = \int_{y\in Y} f\left(y|x \in A\right)\left(y - \hat{y}_t\right)^2 dy.$$

Using

$$\left(y - \hat{y}_t\right)^2 = \left(y - \mu_A\right)^2 + \left(\hat{y}_t - \mu_A\right)^2 - 2\left(y - \mu_A\right)\left(\hat{y}_t - \mu_A\right),$$

we get

$$\begin{aligned} EPE_A\left(C', v^{t-1}\right) &= \int_{y\in Y} f\left(y|x \in A\right)\left(y - \mu_A\right)^2 dy + \int_{y\in Y} f\left(y|x \in A\right)\left(\hat{y}_t - \mu_A\right)^2 dy \\ &\quad - \int_{y\in Y} f\left(y|x \in A\right) 2\left(y - \mu_A\right)\left(\hat{y}_t - \mu_A\right) dy \\ &= Var\left(y_A\right) + \left(\hat{y}_t - \mu_A\right)^2. \end{aligned}$$

Taking expectation over the set of data bases of size $t-1$ we have

$$EPE_A\left(C', t\right) = Var\left(y_A\right) + \mathbb{E}\left[\left(\hat{y}_t - \mu_A\right)^2\right].$$

(ii) If, in categorization C'', A is instead categorized as one separate category then the expected prediction error for objects in A, with expectation taken over the set of data bases of size $t-1$, is

$$EPE_A(C'',t) = Var(y_A)\left(1+\sum_{r=1}^{t-1}\Pr(m_{At}=r)\frac{1}{r}\right)+\Pr(m_{At}=0)\,\mathbb{E}\left[(\hat{y}_t-\mu_A)^2\,|m_{At}=0\right].$$

(iii) Combining (i) and (ii) we have

$$EPE_A(C'',t) - EPE_A(C',t) = Var(y_A)\sum_{r=1}^{t-1}\Pr(m_A=r)\frac{1}{r}$$
$$+\Pr(m_A=0)\,\mathbb{E}\left[(\hat{y}_t-\mu_A)^2\,|m_{At}=0\right]-\mathbb{E}\left[(\hat{y}_t-\mu_A)^2\right].$$

If $\mu_A \neq \mu$ then $E\left[(\hat{y}_t-\mu_{At})^2\right] > 0$ for all t so that for sufficiently large t we have $EPE_A(C'',t) < EPE_A(C',t)$. If instead $\mu_A = \mu$ then

$$\mathbb{E}\left[(\hat{y}_t-\mu_A)^2\right] = Var(\hat{y}_t) = \frac{1}{t-1}Var(y).$$

Also note

$$\mathbb{E}\left[(\hat{y}_t-\mu_A)^2\,|m_{At}=0\right] = Var(\hat{y}_t|m_{At}=0) > Var(\hat{y}_t) = \frac{1}{t-1}Var(y).$$

Thus for $\mu_A = \mu$ we have

$$EPE_A(C'',t) - EPE_A(C',t)$$
$$= Var(y_A)\sum_{r=1}^{t-1}\Pr(m_{At}=r)\frac{1}{r}+\Pr(m_{At}=0)\,\mathbb{E}\left[(\hat{y}_t-\mu_A)^2\,|m_{At}=0\right]-\frac{1}{t-1}Var(y)$$
$$> Var(y_A)\sum_{r=1}^{t-1}\Pr(m_{At}=r)\frac{1}{r}-\left(1-\Pr(m_{At}=0)\right)\frac{1}{t-1}Var(y).$$

We have $Var(y_A) < Var(y)$. Still, as $t\to\infty$ both $\sum_{r=1}^{t-1}\Pr(m_{At}=r)\frac{1}{r}$ and $\frac{1}{t-1}$ go to zero, but the latter does so faster than the former; formally

$$\lim_{t\to\infty}\frac{\frac{1}{t-1}}{\sum_{r=1}^{t-1}\Pr(m_{At}=r)\frac{1}{r}} = \lim_{t\to\infty}\frac{1}{\sum_{r=1}^{t-1}\Pr(m_{At}=r)\frac{t-1}{r}} = 0.$$

It follows that for sufficiently large t we have $EPE_A(C'',t) > EPE_A(C',t)$ when $\mu_A = \mu$. Hence in this case it is better not to categorize A – i.e. have a categorization all of whose categories have an empty intersection with A. These results are about what categorizations that minimize $EPE(C,t)$ for sufficiently large t. Clearly they are readily reformulated as statements about what categorizations that minimize $EPE(C,T,L)$ for sufficiently large L. □

References

Al-Najjar, N. I. and Pai, M. (2009), 'Coarse Decision Making', Manuscript.

Anderson, J. R. (1990), *The Adaptive Character of Thought*, Erlbaum, Hillsdale, NJ.

Anderson, J. R. (1991), 'The Adaptive Nature of Human Categorization', *Psychological Review* 98(3), 409–429.

Ashby, F. G. and Waldron, E. M. (1999), 'On the Nature of Implicit Categorization', *Psychonomic Bulletin and Review* 6, 363–378.

Barberis, N. and Shleifer, A. (2003), 'Style Investing', *Journal of Financial Economics* 68, 161–199.

Berlin, B., Breedlove, D. and Raven, P. (1973), 'General Principles of Classification and Nomenclature in Folk Biology', *American Anthropologist* 74, 214–242.

Bernstein, R. (1995), *Style Investing*, Wiley, New York.

Binmore, K. (2007), 'Making Decisions in Large Worlds', Working paper, University College, London.

Chater, N. (1999), 'The Search for Simplicity: A Fundamental Cognitive Principle?', *Quarterly Journal of Experimental Psychology* 52, 273–302.

Corter, J. E. and Gluck, M. A. (1992), 'Explaining Basic Categories: Feature Predictability and Information', *Psychological Bulletin* 2, 291–303.

Coval, J. D., Jurek, J. and Stafford, E. (2009), 'The Economics of Structured Finance', *Journal of Economic Perspectives* 23(1), 3–25.

Dow, J. (1991), 'Search Decisions with Limited Memory', *Review of Economic Studies* 58, 1–14.

Franklin, A., Clifford, A., Williamson, E. and Davies, I. (2005), 'Color Term Knowledge does not Affect Categorical Perception in Toddlers', *Journal of Experimental Child Psychology* 90, 114–141.

Fryer, R. and Jackson, M. O. (2008), 'A Categorical Model of Cognition and Biased Decision Making', *The B.E. Journal of Theoretical Economics (Contributions)* 8(1), 1–42.

Gärdenfors, P. (2000), *Conceptual Spaces: The Geometry of Thought*, MIT Press, Cambridge, MA.

Gilboa, I., Lieberman, O. and Schmeidler, D. (2006), 'Empirical Similarity', *Review of Economics and Statistics* 88, 433–444.

Gilboa, I., Lieberman, O. and Schmeidler, D. (forthcoming), 'A Similarity-Based Approach to Prediction', *The Journal of Econometrics* .

Gilboa, I., Postlewaite, A. and Schmeidler, D. (2008), 'Probabilities in Economic Modeling', *Journal of Economic Perspectives* 22, 173–188.

Gilboa, I. and Samuelson, L. (2008), 'Preferring Simplicity', Manuscript.

Gilboa, I. and Schmeidler, D. (1995), 'Case-Based Decision Theory', The *Quarterly Journal of Economics* 110, 605–639.

Gilboa, I. and Schmeidler, D. (2003), 'Inductive Inference: An Axiomatic Approach', *Econometrica* 71, 1–26.

Hauser, M., MacNeilage, P. and Ware, M. (1997), 'Numerical Representations in Primates', *Proceeding of the National Academy of the Sciences* 93, 1514–1517.

Herrnstein, R. (1979), 'Acquisition, Generalization, and Discrimination Reversal of a Natural Concept', *Journal of Experimental Psychology: Animal Behavior Processes* 5, 116–129.

Herrnstein, R. J., Loveland, D. H. and Cable, C. (1976), 'Natural Concepts in Pigeons', *Journal of Experimental Psychology: Animal Behavior Processes* 2, 285–302.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999), 'Data Clustering: A Review', *ACM Computing Surveys* 31, 264–323.

Jehiel, P. (2005), 'Analogy-Based Expectation Equilibrium', *Journal of Economic Theory* 123, 81–104.

Jehiel, P. and Samet, D. (2007), 'Valuation Equilibrium', *Theoretical Economics* 2, 163–185.

Johnson, K. E. and Mervis, C. B. (1998), 'Impact of Intuitive Theories on Feature Recruitment throughout the Continuum of Expertise', *Memory and Cognition* 26(2), 382–401.

Jones, G. Y. (1983), 'Identifying Basic Categories', *Psychological Bulletin* 94, 423–428.

Kay, P. and Maffi, L. (1999), 'Color Appearance and the Emergence and Evolution of Basic Color Lexicons', *American Anthropologist* 101(1), 743–760.

Krueger, J. and Clement, R. (1994), 'Memory-Based Judgments About Multiple Categories', *Journal of Personality and Social Psychology* 67, 35–47.

Kuhn, T. S. (1970), *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago.

Laurence, S. and Margolis, E. (1999), 'Concepts and Cognitive Science', in E. Margolis and S. Laurence, eds, *Concepts: Core Readings*, MIT Press, Cambridge, MA, pp. 3–81.

Lipman, B. (2006), 'Why is Language Vague?', Manuscript.

Malt, B. C., Ross, B. H. and Murphy, G. L. (1995), 'Predicting Features for Members of Natural Categories when Categorization is Uncertain', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21, 646–661.

Markman, A. B. andWisniewski, E. J. (1997), 'Similar and Di⊗erent : The Di⊗erentiation of Basic Level Categories', *Journal of Experimental Psvchology: Learning, Memory, and Cognition* 23(1), 54–70.

McKinley, S. C. and Nosofsky, R. M. (1995), 'Investigations of Exemplar and Decision Bound Models in Large, Ill-Define Category Structures', *Journal of Experimental Psychology: Human Perception and Performance* 21, 128–148.

Medin, D. L. (1983), 'Structural Principles of Categorization', in B. Shepp and T. Tighe, eds, *Interaction: Perception, Development and Cognition*, Erlbaum, Hillsdale, NJ, pp. 203–230.

Morris, S. (1995), 'The Common Prior Assumption in Economic Theory', *Economics and Philosophy* 11, 227–253.

Mullainathan, S. (2002), Thinking Through Categories. Mimeo, MIT.

Murphy, G. L. (1982), 'Cue Validity and Levels of Categorization', *Psychological Bulletin* 91, 174–177.

Murphy, G. L. (2002), *The Big Book of Concepts*, MIT Press, Cambridge, MA.

Murphy, G. L. and Ross, B. H. (1994), 'Predictions from Uncertain Categorizations', *Cognitive Psychology* 27, 148–193.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A. and Shafir, E. (1990), 'Category Based Induction', *Psychological Review* 97, 185–200.

Peski, M. (2007), 'Prior Symmetry, Categorization and Similarity-Based Reasoning, Manuscript.

Pothos, E. M. and Chater, N. (2002), 'A Simplicity Principle in Unsupervised Human Categorization', *Cognitive Science* 26, 303–343.

Punj, G. and Moon, J. (2002), 'Positioning Options for Achieving Brand Association: A Psychological Categorization Framework', *Journal of Business Research* 55, 257–283.

Quine, W. V. O. (1969), 'Natural Kinds', in *Ontological Relativity and Other Essays*, Columbia Univ. Press.

Rips, L. J. (1975), 'Inductive Judgments about Natural Categories', *Journal of Verbal Learning and Verbal Behavior* 14, 665–681.

Rosch, E. and Mervis, B. C. (1975), 'Family Resemblances: Studies in the Internal Structure of Categories', *Cognitive Psychology* 7, 573–605.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D. and Boyles-Brian, P. (1976), 'Basic Objects in Natural Categories', *Cognitive Psychology* 8, 382–439.

Rubinstein, A. (1998), *Modeling Bounded Rationality*, MIT Press, Cambridge, MA.

Smith, W. (1965), 'Product Differentiation and Market Segmentation as Alternative Marketing Strategies', *Journal of Marketing* 3-8., 3–8.

Solomon, K., Medin, D. and Lynch, E. (1999), 'Concepts do More than Categorize.', *Trends in Cognitive Science* 3, 99–105.

Tanaka, J. W. and Taylor, M. (1991), 'Object Categories and Expertise: Is the Basic Level in the Eye of the Beholder', *Cognitive Psychology* 23, 457–482.

Uller, C. (1997), *Origins of Numerical Concepts: A Comparative Study of Human Infants and Nonhuman Primates*, MIT Press, Cambridge, MA.

Urban, G. L., Hulland, J. S. and Weinberg, B. D. (1993), 'Premarket Forecasting for New Consumer Durable Goods: Modeling Categorization, Elimination, and Consideration Phenomena', *Journal of Marketing* 57, 47–63.

Warglien, M. and Gärdenfors, P. (2008), 'Semantics, Conceptual Spaces and the Meeting of Minds', Manuscript.

EFI, The Economic Research Institute

Published in the language indicated by the title.
A complete publication list can be found at *www.hhs.se/efi*
Books and dissertations can be ordered from EFI via e-mail: efi.publications@hhs.se

Reports since 2007

2010

Dissertations

Ejenäs, Markus. *Ledning av kunskapsintegration – förutsättningar och hinder : en studie av en fusion mellan IT- och managementkonsulter.*
Öhman, Niclas. *Considering intentions.*

2009

Books

Engwall, Lars. *Mercury meets Minerva: business studies and higher education: the Swedish case.*
Hagberg, Axel. *Bankkrishantering.* Forskning i Fickformat.
Henriksson, Lars. *Marknad eller reglering?: vägval för europeisk telekommunikation.* Forskning i Fickformat.
Holmberg, Carina and Filip Wijkström (eds). *Kunskapsbyggaren: meningsfulla möten och kunskap utan gränser: en vänbok till Åke Danielsson.*
Krohwinkel-Karlsson, Anna. *Oändliga projekt?: Om projektförvaltningens tidsproblematik.* Forskning i Fickformat.
Schriber, Svante. *Att realisera synergier: ledning av värdeskapande vid företagsköp.* Forskning i Fickformat.
Sjöblom, Lisa. *Partner eller kontrollant: en studie av Sidas uppföljning.* EFI Civil Society Reports.
Winberg, Hans, Jon Rognes and Claes-Fredrik Helgesson (eds). *Leading Health Care: organizing healthcare for greater value.*
Östman, Lars. *Towards a general theory of financial control for organisations.*

Dissertations

Almenberg, Johan. *Difficult choices: essays on economic behavior.*
Amado, Cristina. *Four essays on the econometric modelling of volatility and durations.*
Arbin, Katarina. *Individual information system acceptance behavior: an electronic ordering system case.*
Brettell Grip, Anna-Karin. *Funding and accountability: studies of a Swedish and a British chamber orchestra.*
Broback, Anna. *Den värdefulla nöjdheten?: en studie om kundnöjdhet och upplevt värde med kläder över tid.*
Darin, Karin. *Social positions in self-employment: a study of employment structures in artistic production and management consulting.*
Dreber Almenberg, Anna. *Determinants of economic preferences.*
Eriksson Giwa, Sebastian. *Procedural justice, social norms and conflict: human behavior in resource allocation.*
Hasseltoft, Henrik. *Essays on the term structure of interest rates and long-run risks.*
Hellström, Katerina. *Financial accounting quality in a European transition economy: the case of the Czech republic.*
Hernant, Mikael. *Profitability performance of supermarkets: the effects of scale of operation, local market conditions, and conduct on the economic performance of supermarkets.*

Jamal, Mayeda. *Creation of social exclusion in policy and practice.*
Lakomaa, Erik. *The economic psychology of the welfare state.*
Lazareva, Olga. *Labor market outcomes during the Russian transition.*
Lee, Samuel. *Information and control in financial markets.*
Lid Andersson, Lena. *Ledarskapande retorik: Dag Hammarskjöld och FN:s övriga generalsekreterare som scen för karisma, dygder och ledarideal.*
Lindqvist, Göran. *Disentangling clusters: agglomeration and proximity effects.*
Korpi, Martin. *Migration, wage inequality, and the urban hierarchy: empirical studies in international and domestic population movements, wage dispersion and income: Sweden, 1993–2003.*
Kragh, Martin. *Exit and voice dynamics: an empirical study of the Soviet labour market, 1940–1960s.*
Melander, Ola. *Empirical essays on macro-financial linkages.*
Melén, Sara. *New insights on the internationalisation process of SMEs: a study of foreign market knowledge development.*
Murgoci, Agatha. *Essays in mathematical finance.*
Rovira Nordman, Emilia. *Interaction across borders: a study about experiential knowledge development in internationalizing SMEs.*
Salomonsson, Marcus. *Essays in applied game theory.*
Sjöström, Emma. *Shareholder influence on corporate social responsibility.*
Törn, Fredrik. *Challenging consistency: effects of brand-incongruent communications.*
Wennberg, Karl. *Entrepreneurial exit.*
Wetter, Erik. *Patterns of performance in new firms: estimating the effects of absorptive capacity.*
Zubrickas, Robertas. *Essays on contracts and social preferences.*
Åge, Lars-Johan. *Business manoeuvring: a grounded theory of complex selling processes.*

2008

Books

Breman, Anna. *Forskning om filantropi. Varför skänker vi bort pengar?* Forskning i Fickformat.
Einarsson, Torbjörn. *Medlemskapet i den svenska idrottsrörelsen: En studie av medlemmar i fyra idrottsföreningar.* EFI Civil Society Reports.
Helgesson, Claes-Fredrik and Hans Winberg (eds). *Detta borde vårddebatten handla om.*
Jennergren, Peter, Johnny Lind, Walter Schuster and Kenth Skogsvik (eds). *Redovisning i fokus.* EFI:s Årsbok 2008. EFI/Studentlitteratur.
Kraus, Kalle. *Sven eller pengarna? Styrningsdilemman i äldrevården.* Forskning i Fickformat.
Petrelius Karlberg, Pernilla. *Vd under press: om medialiseringen av näringslivets ledare.* Forskning i Fickformat.
Portnoff, Linda. *Musikbranschens styrningsproblematik.* Forskning i Fickformat.
Sjöstrand, Sven-Erik. *Management: från kontorsteknik till lednings- och organisationsteori: utvecklingen på Handelshögskolan under 100 år: 1909–2009.*
Östman, Lars. *Den finansiella styrningens realiteter och fiktioner: de finansiella styrformernas svenska historia, berättelser om Petersson och "Ericsson", finansiell styrning – en ansats till generell teori.*
Östman, Lars. *Mycket hände på vägen från Buchhaltung till Accounting: delar av Handelshögskolan under 100 år.*

Dissertations

Axelson, Mattias. *Enabling knowledge communication between companies: the role of integration mechanisms in product development collaborations.*
Benson, Ilinca. *Organisering av övergångar på arbetsmarknaden: en studie av omställningsprogram.*
Elhouar, Mikael. *Essays on interest rate theory.*
Farooqi Lind, Raana. *On capital structure and debt placement in Swedish companies.*
Granström, Ola. *Aid, drugs, and informality: essays in empirical economics.*

Hvenmark, Johan. *Reconsidering membership: a study of individual members' formal affiliation with democratically governed federations.*
Höglin, Erik. *Inequality in the labor market: insurance, unions, and discrimination.*
Johansson, Marjana. *Engaging resources for cultural events: a performative view.*
Kallenberg, Kristian. *Business at risk. Four studies on operational risk management.*
Kviselius, Niklas Z. *Trust-building and communication in SME internationalization: a study of Swedish-Japanese business relations.*
Landberg, Anders. *New venture creation: resistance, coping and energy.*
Pemer, Frida. *Framgång eller fiasko? En studie av hur konsultprojekt värderas i klientorganisationer.*
Rosengren, Sara. *Facing clutter: on message competition in marketing communication.*
Schilling, Annika. *Kan konsulter fusionera?: en studie av betydelsen av identitet vid en fusion mellan konsultföretag.*
Schriber, Svante. *Ledning av synergirealisering i fusioner och förvärv.*
Sjödin, Henrik. *Tensions of extensions: adverse effects of brand extension within consumer relationship.*
Strandqvist, Kristoffer. *Kritiska år: formativa moment för den svenska flygplansindustrin 1944–1951.*
Strömqvist, Maria. *Hedge funds and international capital flow.*
Söderström, Johan. *Empirical studies in market efficiency.*
Sölvell, Ingela. *Formalization in high-technology ventures.*
Thorsell, Håkan. *The pricing of corporate bonds and determinants of financial structure.*
Ulbrich, Frank. *The adoption of IT-enabled management ideas: insights from shared services in government agencies.*
Östling, Robert. *Bounded rationality and endogenous preferences.*

2007

Books

Andersson, Per, Ulf Essler and Bertil Thorngren (eds). *Beyond mobility.* EFI Yearbook 2007. EFI/Studentlitteratur.
Einarsson, Torbjörn and Filip Wijkström. *Analysmodell för sektorsöverskridande statistik: fallet vård och omsorg.* EFI Civil Society Reports.
Ericsson, Daniel. *Musikmysteriet: organiserade stämningar och motstämningar.*
Samuelson, Lennart (ed). *Bönder och bolsjeviker: den ryska landsbygdens historia 1902–1939.*

Dissertations

Ahlersten, Krister. *Empirical asset pricing and investment strategies.*
Alexius, Susanna. *Regelmotståndarna: om konsten att undkomma regler.*
Andersson, Magnus. *Essays in empirical finance.*
Berg, Bengt Åke. *Volatility, integration and grain bank: studies in harvests, rye prices and institutional development of the parish magasins in Sweden in the 18th and 19th centuries.*
Bianchi, Milo. *Of speculators, migrants and entrepreneurs: essays on the economics of trying your fortune.*
Brodin, Karolina. *Consuming the commercial break: an ethnographic study of the potential audiences for television advertising.*
Elger, Max. *Three essays on investment-specific technical change.*
Hagberg, Axel. *Bankkrishantering: aktörer, marknad och stat.*
Hinnerich, Mia. *Derivatives pricing and term structure modeling.*
Hjalmarson, Hanna. *En växande marknad: studie av nöjdheten med konsumtionsrelaterade livsområden bland unga konsumenter.*
Hjelström, Tomas. *The closed-end investment company premium puzzle: model development and empirical tests on Swedish and British data.*
Kraus, Kalle. *Sven, inter-organisational relationships and control: a case study of domestic care of the elderly.*

Lindqvist, Erik. *Essays on privatization, identity, and political polarization.*
Macquet, Monica. *Partnerskap för hållbar utveckling: systrar av Oikos och guvernanten som blev diplomat.*
Melian, Catharina. *Progressive open source.*
Nilsson, Daniel. *Transactions in cyberspace: the continued use of Internet banking.*
Petrelius Karlberg, Pernilla. *Den medialiserade direktören.*
Portnoff, Linda. *Control, cultural production and consumption: theoretical perspectives, empirical dilemmas, and Swedish music industry practices.*
Sköld, Martin. *Synergirealisering: realisering av produktsynergier efter företagssammanslagningar.*
Sonnerby, Per. *Contract-theoretic analyses of consultants and trade unions.*
Tyrefors, Björn. *Institutions, policy and quasi-experimental evidence.*
Valiente, Pablo. *Re-innovating the existing: a study of wireless IS capabilities to support mobile workforces.*