BETA
Bureau
d'économie
théorique
et appliquée

Documents de travail

« On the Function of Beliefs in Strategic Social Interactions »

Auteur

Arnaud Wolff

Document de Travail n° 2019 – 41

Octobre 2019

**Bureau d'Économie
Théorique et Appliquée
BETA**

www.beta-umr7522.fr

@beta_economics

Contact :
jaoulgrammare@beta-cnrs.unistra.fr

Université de Strasbourg
cnrs
UNIVERSITÉ DE LORRAINE
INRA SCIENCE & IMPACT
AgroParisTech

On the Function of Beliefs in Strategic Social Interactions

Arnaud Wolff

Abstract

We review the way beliefs have traditionally been formalized in game-theoretic settings, and argue that this formalization has its limits, especially in the realm of strategic social interactions. Normative game theory, with its emphasis on equilibrium concepts and its concern about how rational and intelligent players should play, has left little room for a formal characterization of the role of players' beliefs. Given that beliefs determine play, we argue that a case can be made for a deeper understanding of their nature. We draw on the literature in evolutionary psychology and biology to decipher underlying, not readily apparent, incentives that might influence belief adoption. In fact, we take the view that beliefs are themselves subject to incentives, and that agents' beliefs may therefore take on a predictable form if we are able to decipher the underlying incentives that they face. This predictable form might then be used to justify specific modelling assumptions, and accordingly improve the models' predictive power.

Keywords: Beliefs, Game Theory, Social Incentives, Evolution, Coordination

JEL: B40, C70

Bureau d'Economie Théorique et Appliquée (BETA) - Université de Strasbourg - 61 Avenue de la Forêt Noire, 67000 Strasbourg - France. e-mail: arnaudwolff@unistra.fr. I want to thank Jocelyn Donze, Frédéric Koessler, and Gisèle Umbhauer for their useful comments. I am also grateful to Moshe Hoffman, who inspired me to uncover the hidden incentives that pervade our social lives. Several examples in this article are his.

Introduction

> There are many games for which, once the priors are given, the identities of the rational acts follow trivially, and then game theory itself is trivialized if it is merely assumed that the priors are such and such. To avoid this trivialization by Bayesianization, we must take the content of the priors in such cases to be central unknowns of the theory, endogenous to it. (Bacharach and Hurley (1991, p.26))

While game theory is allegedly rooted in Bayesian decision theory (Myerson (1991, p.5)), very little consideration has been given to *Savagian* personal probabilities. In the Savagian framework, beliefs are endogenously determined, together with agent's utility function. There is no presumption about their origin, nor is it required that they be the same across players. In traditional game theory, however, restrictions–based on a particular notion of rationality–are often placed on what the beliefs can be.

The usual argument for restrictions on players' beliefs is the sharpness of the results that ensue. Indeed, it is often posited that if we allow too much subjectivity in players' beliefs, then every outcome is possible, which is true. Aumann (1987, p.15) writes that "[...] the subjective correlated equilibrium is a relatively "weak" concept, giving little information", while Harsanyi and Selten (1988, p.140) argue that strategic play based on different subjective priors "will generally not be an equilibrium point of the game, and therefore it cannot be the outcome chosen by a rational outcome-selection theory". These arguments explain the quasi-universal adoption of the common prior assumption (CPA) in game theory, as well as the recourse to the Harsanyi doctrine.

In the first part of the article, we review in detail how beliefs have traditionally been formalized in game theory. From classical to epistemic game theory, we discuss how restrictions on beliefs gradually fade out, but never completely disappeared. We examine the arguments that have commonly been given for the various restrictions placed on beliefs. Based on the work of Gul (1991) and Morris (1995), we argue that the CPA, that is still widely posited, and the Harsanyi doctrine, that is often called upon to justify the former assumption, do not necessarily hold under scrutiny. Therefore, while these assumptions are needed to have sharp and original results (e.g., the correlated equilibrium as an expression of Bayesian rationality - Aumann (1987)), we find that, in the majority of contexts studied, they do not have much authority. But what happens if we decide to drop these assumptions? Would it require the adoption of solution concepts that are at most "weak" (Aumann (1987, p.15))?

In the second part of the article, we try to defend the idea that a better understanding of the function(s) of beliefs may allow us to answer by the negative to the latter question. Nevertheless, we do not claim to have found a new solution concept, neither do we think that the approach we advocate can resolve every issue related to the formalization of beliefs. Normative game theorists admittedly have a different agenda from the one we try to push forward in this paper. The search for universal, context-free restrictions on (rational) beliefs is a useful theoretical endeavor. We just think that by taking a different perspective, which starts by endogenizing beliefs, we might be able to predict their particular form in the *specific* contexts under investigation, and therefore reap more interesting insights in comparison to a simple *ad hoc* attribution of arbitrary beliefs.

In fact, if beliefs determine play, we argue that a case can be made for a deeper understanding of their nature, and we believe that game-theoretic tools can help us in this endeavor. We start by distinguishing *pragmatic* beliefs, which are the beliefs commonly studied in game theory and economics, from *social* beliefs, which have mainly been neglected, and are not clearly understood. While pragmatic beliefs need to respond to evidence, because the outcome of decisions based on them depends on the true state of nature, social beliefs need not, because they mainly respond to *social incentives*. Based on the literature in evolutionary psychology and biology, we argue that these incentives principally take three forms: the need to appear as a good coalition (group) member, the need to appear as consistent to others, and the need to appear as beneficial and effective to others. We review in detail in exactly what way could these social incentives shape the (mainly social) beliefs individuals adopt, and in exactly what way they could freeze their revision. We also identify gaps in our understanding of how individuals respond to these incentives, and on what the particular trade-offs are. We try to argue that the tools of game theory, which are well-suited to the study of incentive problems, coupled with a better understanding of human motivations, can help us make sense about, for instance, persistent differences in beliefs, and, especially, about what particular form these beliefs will take in different strategic contexts. Therefore, we hope that this approach will reinforce, rather than weaken the insights we might obtain from studying game-theoretic applications.

We proceed as follows. The first part of the article is dedicated to the discussion of how beliefs have traditionally been formalized in game theory. In Section 1.1, we discuss classical and early Bayesian game theory, in which players' beliefs are severely restricted. In Section 1.2, we review arguments for and against the common prior assumption (CPA) and the Harsanyi doctrine. In Section 1.3, we discuss the epistemic program, that places uncertainty about the other players' actions center-stage. In the second part of the article, we discuss the alternative approach that we advocate. Section 2.1 discusses the distinction between *pragmatic* and *social* beliefs, and defines their respective function. In Section 2.2, 2.3 and 2.4 we respectively discuss the motivations (or incentives) to appear as a good coalition member, to appear as consistent to others, and to appear as "beneffective" to others as potential explanations for why individuals adopt the beliefs they do, and why social beliefs might *stick* so much. The last section concludes.

1 On Beliefs in Game Theory

1.1 Classical and (Early) Bayesian Game Theory

Classical Game Theory

Nash formally developed the theory of n-player non-cooperative games. His primary objective was to be able to predict the strategy used by *rational* players. For Nash, a rational prediction should identify *one* solution to the game, the players should be able to correctly identify this solution (and therefore correctly predict the actions of the other players), and this solution should be in every players' best interest (given what is prescribed to the other players). This was his rationale behind his famous concept of strategic (Nash) equilibrium (see Nash (1950, p.23)).

Early theorists have followed Nash's lead by endowing players with a powerful cognition: they are *rational*, in the sense that they "make decisions consistently in pursuit of [their] own objectives" (i.e., maximize their expected utility payoffs), and they are *intelligent*, in the sense that they "know everything that we know about the game, and [...] can make any inferences about the situation that we can make" (Myerson (1991, p.2 and p.4)). Harsanyi and Selten (1988, p.342, emphasis added) write that "the basic task of game theory is to tell us what strategies rational players will follow and *what expectations they can rationally entertain* about other rational players' strategies". Enormous emphasis has therefore been placed on the Nash equilibrium and its refinements, given that strategy profiles that satisfy these equilibrium conditions are the only ones compatible with intelligent and rational play.[1]

The assumptions made about the players necessarily imply that the analysis has to take place at equilibrium. Indeed, (i) a nonequilibrium specification would break down if the players believed it; therefore, (ii) the only way a nonequilibrium specification could be sustained (or prescribed) is if some players have inconsistent beliefs; (iii) given that the players are assumed to be intelligent and rational, their beliefs need to be consistent and correct; (iv) because their beliefs are consistent and correct, equilibrium play is the only possible alternative.

Players' beliefs about other players' actions are accordingly indirectly derived from the equilibrium specification: because the equilibrium strategy profile is the only reasonable one, only beliefs supporting this strategy profile can be justified. This completely restricts the range of beliefs deemed reasonable for a player to hold, to the point of them becoming irrelevant in the analysis. As noted by Brandenburger (2014, p.xx), "Nash banished uncertainty about strategies".[2]

(Early) Bayesian Game Theory

Harsanyi, in a paper about bargaining under ignorance (Harsanyi (1962)), and then in his foundational work on games with incomplete information (Harsanyi (1967), Harsanyi (1968a), Harsanyi (1968b)), was the first to study game situations in which players might be uncertain (i.e., have in mind some probability distribution over a set of parameter values) about some aspects of the structure of the game, such as the other players' utility functions. He developed a methodology (the *type-based approach*) that allowed him to formalize these situations as games with complete (but imperfect) information, and applied to them the concept of Bayesian equilibrium. Harsanyi however did not use his framework to discuss players' uncertainty about the other players' strategies.

In later work, Harsanyi accommodated the case in which players may be uncertain about what the others will play. In Harsanyi (1975), he was, to the best of our knowledge, the first to propose what can be called a *primitive* Bayesian approach, by allowing players to have subjective prior probabilities over the set of strategies

[1] see Kalai and Samet (1984), Kohlberg and Mertens (1986), Kreps and Wilson (1982), Myerson (1978), or Selten (1975).

[2] Another way to look at the Nash equilibrium concept is to see it as a rest point, or as the result of a process of learning or experimentation (c.f. the "steady state" interpretation of the Nash equilibrium, in contrast to the "deductive" one (Osborne and Rubinstein (1994, p.5))). We believe that this interpretation is far more defensible (e.g., see Hoffman et al. (2016) for an insightful discussion about how evolutionary dynamics and the concept of Nash equilibrium can shed light on, among other things, our sense of morality), and stress that it is not the subject of our current discussion.

of the other players. Harsanyi's assumptions were however quite restrictive, given that every player has to share the same given subjective prior probability about the actions of the other players. His *tracing procedure*, which describes how players' expectations about the other players' strategies converge to equilibrium expectations, falls short of rendering players' beliefs anywhere near relevant for the analysis. Whatever the subjective prior beliefs of the players when they enter the game situation, these beliefs will be revised until they all are consistent with the play of an equilibrium strategy profile. As noted by Lecouteux:

> Priors in Harsanyi's model are [...] built such that the outcome of the tracing procedure (i.e., the revisions of players' priors) is a solution of the game. The "Bayesian" dimension of the theory is, therefore, largely artificial since what makes a strategy rational is not that it maximizes one's subjective expected utility *but that it belongs to an equilibrium profile*. (Lecouteux (2018, p.1425, emphasis added))

It is important to note that the restrictive assumptions about players' beliefs are not left undefended, but are backed up by several arguments. The most important argument for why players' prior beliefs might be so constrained (and even identical) is that when the players enter the game situation, all the relevant information is provided by the game structure (the basic parameters of the mathematical model), and given that the players are assumed to be identical, they will all make the same (consistent) inferences about uncertain events. One can therefore assume the existence of an *objective* prior probability distribution from which the players draw their conditional inferences. This view is complemented by the idea that all possible differences in beliefs need to be explained by an asymmetric access to information between the players. These two important ideas, respectively called the *common prior assumption*, and the *Harsanyi doctrine*, are reviewed in the following section.

1.2 The Common Prior Assumption

On the Origins of the Common Prior Assumption

The *common prior assumption* (CPA) refers to the idea, defended by Harsanyi (1967), that it is reasonable to assume that the players in a game all share the same subjective prior probability distribution about uncertain events. It has pervaded game theory in that most theorists assume it in their models. Bernheim (1986) even treats the CPA as a possible axiom for rational choice in strategic environments. The CPA is therefore central, and, as we shall see, important results break down if it is not assumed.

Harsanyi argued that all differences in beliefs must be explained in terms of differences in information (Harsanyi (1968b, p.497); this argument came to be known as the *Harsanyi Doctrine*, a term coined by Aumann (1974, p.92)). This implies that posterior (conditional) probability distributions can be different, because of some asymmetric access to information between the players, but the *basic* (prior) probability distribution needs to be common to all players. The intuition behind this assertion is the following: suppose that two gamblers ascribe different odds of winning to the same horse; gambler A ascribes probability 1/3 to the horse winning, while gambler B ascribes probability 2/3 to the horse winning. One might therefore intuitively consider these beliefs as being the *prior* probability distributions of the two players. However, if we believe in the Harsanyi Doctrine, then these probability distributions are actually *posterior* probability distributions, because the simple fact that they are different necessarily implies that the two players have had access to different information. The true *basic* distribution, from which both gamblers have started, needs to be common between them. This argument justifies the existence of a common probability distribution from which every player draws her inferences during the game situation.

Aumann has repeatedly discussed the Harsanyi Doctrine. In Aumann (1974), he seems rather reluctant to accept it. His 1974 paper was dedicated to the discussion of subjectivity in randomized strategies, in which players can base their choice of strategy on the outcome of *subjective* random devices, that is, "devices on the probabilities of whose outcomes people may disagree" (Aumann (1974, p.67)). Aumann therefore allows players to disagree about the numerical probability associated to some uncertain event (e.g., the outcome of a randomizing device). He even goes on to argue that given the "complex information situation in which the players find themselves" (Aumann (1974, p.94)), assuming differences in subjective probabilities would be valid even if one adheres to the Harsanyi doctrine.

His take will nevertheless be different in Aumann (1976) and in Aumann (1987). In Aumann (1976), he assumes a common prior over the space of states of the world to prove that if an event is common knowledge between the players, then their posterior (conditional) probabilities need to be equal. In Aumann (1987), he shows that if players share a common prior over the set of states of the world, and if it is common knowledge that all players are rational, then "at each state of the world, the distribution of the action n-tuple s is a correlated equilibrium distribution" (Aumann (1987, p.7)). These two important results crucially rely on common priors. Aumann defends his adoption of the CPA in the following way:

> Under the CPA, differences in probabilities express differences in information *only*. Thus the CPA enables one to zero in on purely informational issues in analyzing economic (and other interactive) models with uncertainty. (Aumann (1987, p.14, emphasis in the original))

Rejecting the CPA, and accepting differences in prior probabilities would, Aumann argues, "[yield] results that are far less sharp than those obtained with common priors" (Aumann (1987, p.14)). All that we could obtain are *subjective* correlated equilibrium distributions, which place few restrictions on the possible outcomes. Aumann therefore joins Harsanyi and Selten, who have argued that strategic play based on possibly different subjective priors "will generally not be an equilibrium point of the game, and therefore it cannot be the outcome chosen by a rational outcome-selection theory" (Harsanyi and Selten (1988, p.140)).

Critiques of the Common Prior Assumption

The common prior assumption has always been controversial, and we will now review two important critiques, respectively made by Gul (Gul (1991), Gul (1998)) and Morris (Morris (1995)), that question its authority.

Gul (1991)'s critique is based on the distinction between what he calls the *objective* and the *subjective* information models (OIM and SIM, for short). The term *objective* is "used to refer to situations in which an argument can be made for consensus among agents regarding a particular probability assessment or the mechanism generating the (asymmetric) information" (Gul (1991, p.3)). By contrast, "probabilities that have meaning only as parameters of a model of rational choice behavior (e.g., the Savage model) will be qualified with the term *subjective*" (Gul (1991, p.3, emphasis in the original)).

The OIM describes an actual, physical, statistical experiment. Let X represent a (finite) set of possible parameter realizations, with typical element x. A statistical experiment can be described by a model (I, p) with p a common prior, I $= \{\Omega, (T_i)_{i \in N}, \bar{x}\}$, Ω a set of states of the world with typical element ω, T_i player i's partition of Ω, and $\bar{x}$ a function $\bar{x} : \Omega \rightarrow X$ that associates to each state of the world ω a parameter realization. Crucially, "in an OIM, each state $[\omega \in \Omega]$, and each realization of types [...], corresponds to an actual (physical) contingency" (Gul (1991, p.6)). Given any realization ω, one can construct the players' higher-order beliefs about the value of the parameter $x \in X$, about the beliefs of the other players, their beliefs about other players' beliefs, etc. Hence, infinite hierarchies of beliefs for each player can be constructed following the realization of any state ω. These infinite hierarchies of beliefs correspond to players' posterior beliefs given their information $t_i(\omega)$. In an OIM, it therefore makes sense to talk about a *prior stage*, in which every player shares the same common prior probability distribution p, and after which players update their beliefs (at the interim stage) given the information they receive about the realization of the true (physical) state of the world.

The SIM is derived differently. One does not start with an actual, physical, statistical experiment, but with players' infinite hierarchies of beliefs describing their beliefs about X, their beliefs about each other's beliefs, etc. Mertens and Zamir (1985) and Brandenburger and Dekel (1993) have shown that, starting with any such infinite hierarchies of beliefs, one can construct some model $(\omega, I, (p_i)_{i \in N})$ with ω a state of the world, I $= \{\Omega, (T_i)_{i \in N}, \bar{x}\}$, and $(p_i)_{i \in N}$ a subjective prior probability *for each player*. There is no common prior for the players, because there is no actual *ex ante* stage at which players' information is symmetric. In a SIM, players' probabilities p_i are not beliefs in the Bayesian sense, but only mathematical constructs. Players' beliefs are represented by the collection $(\omega, I, (p_i)_{i \in N})$. Gul therefore argues that in the SIM, the CPA is "meaningless rather than unsound" (Gul (1991, p.8)). Given the limitations of the OIM with respect to the scope of its possible applications, researchers (Aumann in particular) have mainly been working with the SIM. The difficulties raised by Gul seem serious enough for us to remain doubtful about the relevance and the applicability of the CPA.

Morris (1995) has a more philosophical critique of the CPA. He first argues, as does Gul (1991), that it only makes sense to assume a common prior if there exist *ex ante* objective probabilities to which the players' subjective probabilities could be equal. He writes that "there are special problems defending the common prior assumption in a model where beliefs are endogenously determined" (Morris (1995, p.234)).

With respect to the Harsanyi doctrine, Morris writes that even if a logical relation between information and beliefs existed (an argument that is at the core of the Harsanyi doctrine), there are some situations in which its application would be vain (Morris (1995, p.236)). Without a deeper understanding about the process of belief formation, one might unfortunately be at risk of indulging in circular reasoning.

Concerning the assertion that common priors are justified because with enough learning and experimentation, differences in beliefs will necessarily fade out, Morris notes that the conditions under which learning needs to take place for the process to converge to the true parameter value are very restrictive (Morris (1995, p.237-8)). Most importantly, it is well known that learning and experimentation are not free of biases (see, for instance, Epley and Gilovich (2016) and Kunda (1990)). Therefore, if players don't have accuracy motives, there is no reason why we should expect their beliefs to converge in the long run.

Finally, what are the prospects in terms of prediction if we drop the CPA? Is it true that anything can be rationalized under heterogeneous priors? One can indeed rationalize a lot of results by positing *ad hoc* heterogeneous priors. The key lies in justifying the form players' beliefs take. If one can justify why players hold the beliefs they do, then the *post hoc* rationalization argument does not hold anymore.

1.3 The Epistemic "Revolution"

The epistemic program, arguably launched by Bernheim (1984) and Pearce (1984), has the objective to break away from the strongest assumptions of classical and early Bayesian game theory, by allowing players in a game situation to be uncertain about the other players' actions, as well as by not *requiring* players to share a common prior about all relevant uncertainty in the game. Another objective of the epistemic enterprise is to develop a mathematical framework that allows researchers to be able to make precise formal statements about the players' epistemic states.[3]

Epistemic game theory sets players' uncertainty about what the others will play center-stage, but it appears that this uncertainty, as in early Bayesian game theory, is not unrestricted. As noted by Perea (2014, p.11, emphasis in the original), "a major task of epistemic game theory is to put some *plausible restrictions* on such belief hierarchies, as to distinguish *reasonable* from *less reasonable* belief hierarchies". So, while Brandenburger notes that epistemic game theory can accommodate the case in which players are irrational, or believe that others are irrational (Brandenburger (2010)), epistemic game theory has mainly been concerned with belief hierarchies that are *consistent*, in the sense that every player is rational, believes that every other player is rational, believes that every other player believes that every player is rational, and so on. This is commonly called *common knowledge (or belief) in rationality*. This concern is clear in the following quote from Perea:

[3] The *state space* (or *semantic*) representation of beliefs and knowledge is principally used to investigate issues related to the knowledge and beliefs of players in epistemic game theory (see Battigalli and Bonanno (1999)).

> Since most other concepts in epistemic game theory can be seen as some sharpening, or variant, of common belief in rationality, we may indeed say that the concept of common belief in rationality is the cornerstone of epistemic game theory. (Perea (2014, p.13))

Consistency therefore requires players not to attribute positive probability to any strategy deemed unreasonable for another (rational) player to play. This requirement seems straightforward, because as noted by Bernheim (1986, p.477), "no player will ever choose a dominated strategy, so we can delete all such strategies without changing the game in a substantive way". But as noted by Lecouteux (2018, p.1434), assuming common belief in rationality is actually equivalent to assuming equilibrium play. This stems from Brandenburger and Dekel (1987), who showed that rationalizable strategies are actually strategies that constitute an *a posteriori equilibrium* of the game, which is a refinement of the subjective correlated equilibrium from Aumann (1974). This implies that by assuming common belief in rationality, one only accepts beliefs that are consistent with *some kind* of equilibrium play. As a consequence, "the beliefs of the players are not the primitive of the analysis, but are defined *ex post* in order to evaluate various solution concepts" (Lecouteux (2018, p.1442, emphasis in the original)).

The epistemic program, whose primary objective was to "make epistemic states of players an input of the game" (Brandenburger (2010, p.65)), has not completely dealt with the major issues imputed to the classical approach. Together with the assumption of common priors, the assumption of common belief in rationality removes a crucial part of subjectivity in players' beliefs, thereby weakening their relevance in the theory of games. Most importantly, the restrictions placed on beliefs are somewhat arbitrary, based on a particular (i.e., normative) notion of rational beliefs. Instead of universal restrictions on beliefs, we would like to argue, in the second part of the paper, for a more *case-based* analysis of beliefs, grounded in the idea that different strategic situations will influence players' beliefs in different ways.

Discussion

We have seen, in the first part of this paper, that when strategic interactions are described using the tools of game theory, what players believe or conjecture about their physical and social environment plays a crucial role in the way the game ensues. As noted in the introducing quote from Bacharach and Hurley, "there are many games for which, once the priors are given, the identities of the rational acts follow trivially" (Bacharach and Hurley (1991, p.26)). We also have argued that the way the players' beliefs have traditionally been formalized has severe limitations. Restrictions have often been placed on the range of beliefs players might entertain, but those restrictions are often based on a normative conception of rationality that might not always be relevant. As will be discussed in more detail in Section 2, an important distinction has to be made between *normative* and *ecological* rationality;[4] this distinction will be crucial to garner greater insights into the observed patterns of belief adoption.

[4] For a definition of ecological rationality, see Tooby et al. (2006, p.104-105). For a discussion on the application of inappropriate normative standards on human cognition, see Haselton et al. (2015).

If beliefs determine play, a case can be made for a deeper understanding of their nature. It might not be enough to attribute to players *ad hoc* or arbitrary priors. If we want, for instance, to have deeper insights about the nature of disagreement, it will be necessary to understand the *process* of belief formation. Positing heterogeneous prior beliefs is not sufficient; it is precisely those heterogeneous prior beliefs that need explaining.

In the second part of this paper, our main objective will be to try to convince the reader that the tools of game theory can, if properly applied, help us make sense about the wide distribution of beliefs among individuals (or groups), and can therefore be employed to justify the way those beliefs are formalized in specific strategic interactions.

2 What are Beliefs for?

> We should not assume that human minds are designed to acquire true information about their natural and social environments. (Boyer (2018, p.70))

It will be argued, in the second part of this article, that a better understanding of the function of beliefs might enable researchers to make better predictions about the form they take, especially about how they themselves are influenced by the strategic situation. This approach will be *functional*, influenced by the literature in evolutionary psychology and biology, in that we will examine in detail the specific role(s) of beliefs, particularly what factors might shape them. As a consequence, we take the view that beliefs are themselves subject to incentives, and that agents' beliefs may therefore take on a *predictable* form if we are able to decipher the underlying incentives that agents face. This predictable form might then be used to justify specific modelling assumptions, and accordingly improve the models' predictive power.

2.1 On the Function(s) of Beliefs

It may be argued that our beliefs can be divided in two (broad) functional categories, that might of course overlap. Both functions are evidently linked to survival, but they differ conceptually. We will call the first kind of beliefs *pragmatic*, and the second kind *social*.

Pragmatic beliefs are supposed to depict a reliable map of our physical and social environment. They are most useful in games against Nature. Nature is cold and ruthless, and can not be fooled. Pragmatic beliefs therefore need to keep track of reality, and be updated whenever new evidence arrives. For instance, one needs to have a reliable representation of traffic before crossing the street, or have a consistent assessment of the risks of climbing Mount Everest. Pragmatic beliefs therefore need to be closest to truth, because, intuitively, "true beliefs aid in accomplishing goals and, with appropriate inference machines, generating additional true beliefs" (Kurzban and Christner (2011, p.285)). These are the type of beliefs traditionally studied in economics, game theory, or decision theory; they need to be updated using Bayes' rule (under the appropriate conditions), and they need to be as close as possible to the true value of the parameter of interest.[5] Hence, implicitly assumed is the idea that the agents in our models *want* to get closer to truth, and that they therefore undertake (sometimes) costly actions (such as information acquisition) to accomplish their objective.

Social beliefs are of a different kind. They do not respond to evidence the way pragmatic beliefs do, and the reason is that they actually *shouldn't*. Social beliefs are far less well understood, but probably as much important as pragmatic ones. A general idea about their function is given in the following quote by Pinker:

> People are embraced or condemned according to their beliefs, so one function of the mind may be to hold beliefs that bring the belief-holder the greatest number of allies, protectors, or disciples, *rather than beliefs that are most likely to be true*. (Pinker (2005, p.18, emphasis added))

While pragmatic beliefs are useful in our games against Nature, social beliefs are useful in our games with Others. One of their principal particularity is that we do not have to make decisions based on them, such as betting, whose outcomes depend on the true state of the world. They instead mainly respond to *social incentives*. Their value lies in how they are perceived by others, on the kind of inferences others make about us depending on the beliefs we hold.

Social beliefs have not received the same attention as pragmatic beliefs; maybe this reflects an overall tendency to treat agents in our models as rational truth-seekers. Nevertheless, it can be argued that game-theoretic tools are particularly well-suited to their study. The key lies in understanding what *motivates* people to hold certain beliefs, and what precisely the trade-offs are. What are the principal underlying social (or pecuniary) incentives that regulate the adoption of social beliefs? In exactly what sense can it be advantageous to be wrong about our physical and social environment? Are social beliefs mere signalling devices, independent of our everyday actions, or do they actually constrain our behavior? Why are social beliefs often shared at the level of a group, and why do different groups (or coalitions) share different beliefs?

Answering these questions will be of particular importance if we want to have a better understanding of why humans believe the (sometimes seemingly indefensible) things they do. We will try to defend the idea that the tools of game theory, coupled with a deeper appreciation of human motivations, can help us make sense about patterns of belief adoption. Important work in the areas of evolutionary psychology and biology has already paved the way for a greater understanding of human nature and motivations, and can therefore provide us with hints about where exactly we should look to uncover the incentives that guide our behavior in general, and beliefs in particular.

We will review three, principally hidden motivations that seem, to us, important in explaining the adoption of specific (mainly social) beliefs. All three motivations are obviously not mutually exclusive; they probably act in concert. These are the motivation to appear as a good coalition (or group) member, the need to appear as consistent, and the need to appear as beneficent and effective to others.[6] We will argue that the adoption of a

[5] In these particular cases, it therefore makes sense to study rational (i.e., Bayesian) beliefs, and argue that players *should* be using Bayes' rule to update their beliefs; what one might call *normative* rationality.

[6] Or "Beneffective", as Robert Trivers would put it.

game-theoretic perspective can, in all three cases, improve our understanding of these issues, and illuminate how strategic social life is from an evolutionary perspective.

2.2 Groups In Mind

Humans have evolved in (relatively) small groups of nomadic hunter-gatherers. While it was long believed that these forager bands were mainly comprised of close kin (genetically related individuals), recent evidence suggests that the bands' composition was far more diverse. Indeed, studies of contemporary hunter-gatherer communities have shown that "primary kin make up less than 10% of a residential band" (Hill et al. (2011, p.1288)). Therefore, it is now believed that early humans have evolved in an environment in which a strong selection pressure was to be able to achieve successful coordination with mainly non-genetically related individuals, in order to survive in harsh environments characterized principally by the need to find food and shelter, and to resist pressures and attacks from other groups (see Bowles (2009)).

The fact that we have lived in relatively small bands for most of our species' history suggests that we have evolved adaptations for group-living (i.e., adaptations that allow us to better navigate the social world). This is the thesis defended by, among others, leading evolutionary psychologists Leda Cosmides and John Tooby (see, for instance, Kurzban et al. (2001) Tooby et al. (2006) or Tooby and Cosmides (2010)). The argument is that humans have evolved neural programs for efficient coalitional management, given that the formation and management of coalitions was an evolutionary imperative in ancestral environments.

What is of interest to us is that these cognitive adaptations may underlie our adoption of particular *social* beliefs.[7] Indeed, we argue that by taking this idea seriously, we might be able to uncover important incentives that are not immediately apparent, but that could help us explain why we don't observe absolute convergence of beliefs on largely factual matters.

If humans have adaptations for coalitional formation or affiliation, then one way one can commit to a certain group is by adopting the beliefs held by this particular group. In fact, "to earn membership in a group, you must send signals that clearly indicate that you differentially support it, compared to rival groups" (Tooby (2017)). Motivations for joining groups, and for being seen as a good coalition member, might therefore cause our brains to automatically and unconsciously select the appropriate beliefs.[8] Moreover, by adopting the beliefs of a certain group, one often dissociates from other groups (think about the issues of gun control, abortion, or climate change that mainly oppose Liberals and Republicans in the U.S.), thereby increasing other coalition members' trust toward oneself. To get a concrete example of this general idea, consider the following excerpt from Hochschild, in which she interviews a Christian woman from the United States:[9]

[7] Before we proceed, it is crucial to note that none of these processes need to be conscious. Boyer notes that "because we evolved as support seekers, and therefore recruitment specialists, we can orient our behavior toward more efficient coordination with others without having to be aware of it" (Boyer (2018, p.85)).

[8] For evidence that our unconscious plays an integral and important part in our everyday life, see Von Hippel and Trivers (2011).

[9] See also McCullough et al. (2016) for similar evidence from laboratory experiments.

> "If I know a person is a Christian," one woman told me, "I know we have a lot in common. I'm more likely to trust that he or she is a moral person than I would a non-Christian." (Hochschild (2018, p.217))

It follows that one potential incentive for adopting particular beliefs is the human motivation to seek social support, to join groups, and to be seen as a trustworthy and reliable group member. While this is at least logically valid, little is known about how exactly these incentives operate. As noted by Boyer, "if that is a function of coalitional signaling, we should expect highly stable coalitions to favor commitment signals that are irreversible" (Boyer (2018, p.51)). What are the conditions on beliefs for them to be credible signals of commitment? How can beliefs signal commitment when they are by definition not observable and easily revised? What are the exact trade-offs individuals face? Under what conditions is belonging to one group detrimental to membership in another? What determines which group (and therefore which beliefs) individuals will join (adopt)? What factors can help us explain individuals moving from one group to another?

The point we want to make is twofold. First, a better understanding of human motivations can help us uncover trade-offs that we would not even have considered otherwise. If we only contemplate humans as rational, intelligent, individualistic truth-seekers, then we completely miss potentially crucial (hidden) incentives. Second, once we acknowledge the existence of these otherwise neglected motivations, several new (above-mentioned) questions and interrogations arise. We believe that game-theoretic tools can be applied to, at least partially, answer these questions. This belief is based on the observation that, at the core, the issue is one of trust and coordination. We, humans, need to be able to coordinate with one another, and to coordinate successfully, we need to send each other signals, trust these signals, and coordinate on these signals (Tomasello et al. (2012)). We conjecture that particular beliefs might be adopted to better coordinate with other group members, but that this, by itself, decreases coordination opportunities with non-members, by increasing uncertainty about play; an idea that might by formalized using coordination games (c.f. Kets and Sandroni (2017)). Once one has this framework, more detailed and more difficult questions can be addressed.

2.3 On the Need to Appear Consistent (And the Implications for Theories of Persuasion)

Individuals strive to avoid what has been termed cognitive dissonance (Festinger (1957)), and therefore seek cognitive consistency. They allegedly want that their beliefs be consistent one with another, and in accordance with their everyday behavior. Several explanations have been given for why that would be the case. Having inconsistent beliefs presumably feels bad (but why?), so people want to avoid it. Also, acting in a way that is not consistent with one's preferences or beliefs makes people uncomfortable, so they rationalize this behavior by purportedly adjusting their beliefs and preferences (in order to keep intact their *self-image*).

This all appears as intuitive, but we would rather conjecture that the principal mechanism (or incentive) behind this dissonance avoidance is to appear consistent *to others*, not to oneself. If we take the evolutionary perspective seriously, then it would not make much sense to equip our minds with a "dissonance reduction" mechanism just for its own sake, to make us feel better or more comfortable. If beliefs are in contradiction one with another, or if beliefs are inconsistent with observed reality, then this by itself is useful information, and we should act on it; we should not strive to find *ad hoc* auxiliary hypotheses that make the relationship between our beliefs and our behavior (at least *plausibly*) consistent. The fact that we often indulge in such behavior may be a sign that there are some other underlying motives at play, such as the motivation to persuade others, or to appear as consistent to others. In line with this idea, numerous studies on energy or water conservation have shown that public commitments are more effective than private commitments in inducing change in behavior (Abrahamse and Steg (2013), Lokhorst et al. (2013), Pallak et al. (1980)). This suggests that it is not inconsistency *per se* that matters, but rather the *appearance* of inconsistency.

If people strive to appear consistent to others, then we might deduce that one possible (effective) way to persuade someone about something is to present her with evidence that is inconsistent with a belief she dearly holds. In order not to appear as a fool to others, she will want to modify her prior belief, in order to make it consistent with observed reality. But this would be too easy. In fact, the existence of ad hoc auxiliary hypotheses,[10] that can make almost every argument look *plausible* (defensible) on the surface, is a severe impediment in matters of persuasion. What the above argument suggests is that if you want to change the worldview of an individual, it will not be enough to try to convince her about a single fact, but you would need to address all possible ad hoc auxiliary hypotheses she might call upon. Only when one can not *plausibly* defend one's worldview anymore, and therefore can not possibly reconciliate one's belief with observed reality, will she concede.[11] For instance, Galperti (2019, p.997) has a passage in which he models an individual, Susan (who believes that human activities are likely to cause global warming), who wants to persuade Rick (who denies this as impossible) about that fact. Leaving aside the issue that this approach does not consider the likely underlying incentives to hold these particular beliefs, it is very likely that if Susan shows Rick extremely convincing evidence that human activities affect global warming, then Rick might call on an ad hoc auxiliary hypothesis that is specifically tailored to accommodate disconfirmatory evidence.[12]

In fact, Gershman (2019) has shown that with recourse to ad hoc auxiliary hypotheses, one might possibly never be wrong. With the appropriate prior beliefs, it might even be rational (in the Bayesian sense) to update one's beliefs in this way. Indeed, Gershman writes:

> A large body of empirical work on belief polarization was interpreted by many social psychologists as evidence of irrational belief updating [...]. However, another possibility is that belief polarization might

[10] As noted by Gershman (2019, p.13), "an auxiliary assumption becomes an ad hoc hypothesis when it entails unconfirmed claims that are specifically designed to accommodate disconfirmatory evidence."

[11] Festinger's study of the millennial cults has shown that even in the face of extreme disconfirmatory evidence (that the world did not in fact end), people were still capable of finding excuses and rationalizing the non-event. This implies that it may be very unlikely that we might be able to persuade people on matters on which they have no incentives to be right.

[12] For instance, that scientists can not be trusted on these matters, that the medias are making an unwarranted fuss about it, or that it is sometimes still cold and snowing.

> arise from different auxiliary hypotheses about the data-generating process [...]. If participants assume the existence of research bias (distortion or selective reporting of findings to support a preconceived conclusion), then reading a study about the ineffectiveness of the death penalty may strengthen their belief in research bias, correspondingly increasing their belief in the effectiveness of the death penalty. (Gershman (2019, p.19))

Ad hoc auxiliary hypothesis, which are often called upon to maintain a plausibly consistent (defensible) worldview, are therefore a clever tool one might use to avoid having to change one's beliefs. It follows that without a better understanding of the motivations driving the adoption of specific beliefs, efforts to persuade will likely not be successful. Before trying to persuade someone (or some group), we need to understand how they came to adopt their belief (which will often be explained, we conjecture, by underlying incentives rather than asymmetric information access). As noted by Simler and Hanson (2017, p.311), "it is only by understanding where the resistance is coming from that we have any hope of overcoming it".

2.4 On the Need to Appear "Beneffective"

Why are we often overconfident (Barber and Odean (2001), Glaser and Weber (2007)), or believe that we have more control than we actually have (Fenton-O'Creevy et al. (2003), Langer and Roth (1975))? Why do we often stay optimistic, even in the face of bad news, and why are we reluctant to accept negative information about ourselves, but update consistently when the news are good (Eil and Rao (2011))? Why is it that we sometimes avoid getting information about potentially deadly diseases that we might have (Sweeny et al. (2010))? Why do we misremember, or forget, negative information about ourselves (Croyle et al. (2006)), but have no problem in remembering positive information (Green et al. (2008))? Why do we think that we are better (on a large range of domains) than we actually are (Epley and Whitchurch (2008), Guenther and Alicke (2010))?

We might say that people are irrational, or that they have cognitive biases. We might also say that they only indulge in wishful thinking: they want to feel good about themselves, so they adopt beliefs that are in accordance with what they would like the world to be. They believe what they want to be true. But that is just too easy (and very unlikely from an evolutionary perspective). It stops the investigation just where it should start.

From the outside, it appears as if people *self-deceive*. They overestimate their own qualities, and rationalize away negative information (even sometimes avoid it). Why might this be? Shouldn't our brains have evolved in order to present an accurate description of the world and our abilities, and take every possible bit of information into account (including the fact that everyone thinks that they are better than others, which is by itself useful information)? Standard decision theory has shown that in single-decision problems (games against Nature), having more information never hurts the decision-maker (Osborne and Rubinstein (1994, p.71)). However, when we enter the realm of strategic interactions (games against Others), having less information, or at least *appear* to have less information, can be very useful to the decision-maker (Bassan et al. (1997)). The key to this puzzle (to why people distort or avoid information) may therefore lie in underlying strategic incentives.

In fact, a common theme among all these behaviors is that people are trying to exhibit the most positive, defensible, image of themselves to others. People do not deceive themselves just to feel good; they deceive themselves to persuade others about their qualities. This is the principal argument made by Kurzban (2011) and Trivers (2011). By strategically self-deceiving, we all play persuasion games one with another. We don't attend to potentially detrimental information (or strategically forget about it), so that if others query us about it, we can safely respond that we were unaware, thereby avoiding lying.[13] We (unconsciously) enhance our self-image, and we internalize this belief, not to feel better about ourselves, but to more easily convince others about our worth as a friend or a mate. We are overconfident, and feel that we have control, so that we may persuade others that we are in charge, and that we know what we are doing.

In work on motivated reasoning, it is often stated that people distort their beliefs in the direction of what they would like to be true, or in such a way that the beliefs they hold make them feel good. For instance, Bénabou and Tirole write:

> Some beliefs and emotions are affectively more pleasant than others, like hope and confidence over fear and anxiety. People receive utility from having a positive self-image, and from thinking of themselves as belonging to groups. Optimistic beliefs can also be valuable motivators to overcome self-control problems, as well as helpful in strategic interactions. (Bénabou and Tirole (2016, p.160-1))

Bénabou and Tirole acknowledge that misrepresentations can have advantages in strategic interactions, but it seems that they consider this feature as a by-product, rather than as the *actual* function of the misrepresentation. Moreover, by stopping the analysis at the *proximate* level (i.e., at the level of conscious thought), we might possibly miss the *ultimate* function of these particular beliefs (i.e., *why* they feel good or bad).[14]

The two approaches - self-deception for one's own good, and self-deception to deceive others - often overlap, but they are conceptually very different. The predictions one might make under one approach might be considerably different than the predictions made under another. As a matter of example, if we did believe only what we wanted to be true (or what feels good), but not what we wanted to persuade others of, then far-right conservatives would not believe that migrants are looters or murderers, and extreme feminist movements would not be convinced about the oppression of women by men.

Discussion

The key point we want to underline is that in strategic interactions, we might be able to predict the direction of people's beliefs based on the underlying incentives that they face. Nonetheless, it is important to remember that all these incentives probably act in concert. For instance, one might want to be seen as acting consistently with respect

[13] See DePaulo et al. (2003) for evidence that when telling lies, people often get spotted.

[14] For a discussion of the distinction between proximate and ultimate causation, see Scott-Phillips et al. (2011).

to a particular (core) belief adopted by one's own group. That is, the core belief, around which other related beliefs are tailored, is endogenously determined by one's group belonging. Also, one might want to stay consistent with respect to a cherished belief, in order to appear as effective to others. A politician, for instance, might want to stay true to her principles in order to appear as confident, or in charge, thereby persuading others of her convictions.

The primary task of the theorist should be to decipher which incentives are the most relevant in the particular context studied (and why?). In fact, very little is known about the interplay of the different incentives individuals face, and about which ones are more binding (and when?). Pecuniary incentives might well be as important as social incentives (think about oil company managers that are convinced that their activities are not harmful for the environment). When, for instance, do pecuniary incentives take the ascendant over social incentives? Or, conversely, in which cases do social incentives (for instance, the incentive to appear as a good coalition member), trump financial incentives? It will likely be very important to determine how individuals deal with the numerous (hidden, underlying) trade-offs that they face, if we want to reap more interesting insights about human behavior. But to be successful, this endeavor requires that we take more seriously the evolutionary approach to human motivations.

Conclusion

Aumann (1976) has shown that if individuals start with the same priors beliefs, and their posteriors about the occurrence of some event are common knowledge, then these posteriors must be equal (i.e., their beliefs must be the same). One however often observes persistent disagreement. Is it because the respective posteriors only rarely fully become common knowledge, or is it because people don't start off with the same priors? Geanakoplos and Polemarchakis (1982) have shown that in a finite period of time, *honest, truth-seeking* individuals will reach an agreement by communicating back and forth their posteriors, even if the event was not common knowledge to start with. It must therefore be the priors. Or is it because people are not honest, truth-seekers? We believe that the answer is a mix of both, and that the key to understanding why people are not honest, truth-seekers, and therefore do not converge towards a common posterior, is to decipher the underlying incentives they face to adopt their respective beliefs.

It has traditionally been assumed in economics and game theory that the agents in our models want to be *right*. They strive to get closer to truth, and they will sometimes undertake costly actions to reach their objective. This assumption is evidently uncontroversial when we consider agents taking decisions whose outcomes depend on the true state of nature. A trader investing in the stock market will want the best available information, while environmentally-friendly consumers will want to know everything about the products they buy. The beliefs need to be *pragmatic*, grounded in reality. But social life is complex, and there are other (maybe even more important) incentives influencing our beliefs and behavior. We have argued, based on the literature in evolutionary psychology and biology, that some additional, not readily apparent motivations presumably play a large role. These motivations comprise the need to appear as a good coalition (group) member, the need to appear as consistent, and the need to appear as beneficent and effective to others. These motivations are distinct (and without any doubt not exhaustive), but they probably act in concert. In all these cases, pragmatic beliefs are not very useful, because agents have no incentives to be right. Their decisions in these areas do not directly depend on the true state of nature. Instead, their incentives are social, and their beliefs will bear this sign.

Social beliefs have not been much studied, and they are not well understood. They do not respond to evidence as pragmatic beliefs do, and the reason is that they shouldn't. We have conjectured that much of the apparent disagreement on largely factual matters is due to the above mentioned motivations (together with pecuniary incentives), and that without a deeper appreciation of the underlying (hidden) incentives that agents face, we will not be able to improve our understanding of how to tackle pressing issues such as climate denial, conspiracy theories, or anti-science movements. We believe that the tools of game theory can be successfully applied in those areas, helping us decipher which incentives are the more stringent and binding in different contexts. Nonetheless, this new endeavor requires a deeper comprehension of human motivations and of the particular (often hidden) incentives that we face, at the risk of being stuck in the study of *proximate* mechanisms, without grasping what the *ultimate* motives are.

References

Abrahamse, W. and Steg, L. (2013). Social influence approaches to encourage resource conservation: A meta-analysis. *Global environmental change*, 23(6):1773–1785.

Aumann, R. J. (1974). Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics*, 1(1):67–96.

Aumann, R. J. (1976). Agreeing to disagree. *The annals of statistics*, pages 1236–1239.

Aumann, R. J. (1987). Correlated equilibrium as an expression of bayesian rationality. *Econometrica: Journal of the Econometric Society*, pages 1–18.

Bacharach, M. and Hurley, S. (1991). Issues and advances in the foundations of decision theory. *Foundations of decision theory*, pages 1–38.

Barber, B. M. and Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The quarterly journal of economics*, 116(1):261–292.

Bassan, B., Scarsini, M., and Zamir, S. (1997). "i don't want to know!": Can it be rational?. hebrew university jerusalem. *Center for Rationality and Interactive Decision Theory, Discussion Paper*, 158.

Battigalli, P. and Bonanno, G. (1999). Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics*, 53(2):149–225.

Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–64.

Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica: Journal of the Econometric Society*, pages 1007–1028.

Bernheim, B. D. (1986). Axiomatic characterizations of rational choice in strategic environments. *The scandinavian journal of economics*, pages 473–488.

Bowles, S. (2009). Did warfare among ancestral hunter-gatherers affect the evolution of human social behaviors? *Science*, 324(5932):1293–1298.

Boyer, P. (2018). *Minds make societies: How cognition explains the world humans create*. Yale University Press.

Brandenburger, A. (2010). Origins of epistemic game theory. *Epistemic logic: Five questions*, pages 59–69.

Brandenburger, A. (2014). *The language of game theory: Putting epistemics into the mathematics of games*, volume 5. World scientific.

Brandenburger, A. and Dekel, E. (1987). Rationalizability and correlated equilibria. *Econometrica: Journal of the Econometric Society*, pages 1391–1402.

Brandenburger, A. and Dekel, E. (1993). Hierarchies of beliefs and common knowledge. *Journal of Economic Theory*, 59(1):189–198.

Croyle, R. T., Loftus, E. F., Barger, S. D., Sun, Y.-C., Hart, M., and Gettig, J. (2006). How well do people recall risk factor test results? accuracy and bias among cholesterol screening participants. *Health Psychology*, 25(3):425.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., and Cooper, H. (2003). Cues to deception. *Psychological bulletin*, 129(1):74.

Eil, D. and Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–38.

Epley, N. and Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic Perspectives*, 30(3):133–40.

Epley, N. and Whitchurch, E. (2008). Mirror, mirror on the wall: Enhancement in self-recognition. *Personality and Social Psychology Bulletin*, 34(9):1159–1170.

Fenton-O'Creevy, M., Nicholson, N., Soane, E., and Willman, P. (2003). Trading on illusions: Unrealistic perceptions of control and trading performance. *Journal of occupational and organizational psychology*, 76(1):53–68.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford university press.

Galperti, S. (2019). Persuasion: The art of changing worldviews. *American Economic Review*, 109(3):996–1031.

Geanakoplos, J. D. and Polemarchakis, H. M. (1982). We can't disagree forever. *Journal of Economic theory*, 28(1):192–200.

Gershman, S. J. (2019). How to never be wrong. *Psychonomic bulletin & review*, 26(1):13–28.

Glaser, M. and Weber, M. (2007). Overconfidence and trading volume. *The Geneva Risk and Insurance Review*, 32(1):1–36.

Green, J. D., Sedikides, C., and Gregg, A. P. (2008). Forgotten but not gone: The recall and recognition of self-threatening memories. *Journal of Experimental Social Psychology*, 44(3):547–561.

Guenther, C. L. and Alicke, M. D. (2010). Deconstructing the better-than-average effect. *Journal of Personality and Social Psychology*, 99(5):755.

Gul, F. (1991). *On the Bayesian View in Game Theory and Economics*. Graduate School of Business, Stanford University.

Gul, F. (1998). A comment on aumann's bayesian view. *Econometrica*, 66(4):923–927.

Harsanyi, J. C. (1962). Bargaining in ignorance of the opponent's utility function. *Journal of Conflict Resolution*, 6(1):29–38.

Harsanyi, J. C. (1967). Games with incomplete information played by bayesian players, i–iii part i. the basic model. *Management science*, 14(3):159–182.

Harsanyi, J. C. (1968a). Games with incomplete information played by bayesian players part ii. bayesian equilibrium points. *Management Science*, 14(5):320–334.

Harsanyi, J. C. (1968b). Games with incomplete information played by bayesian players, part iii. the basic probability distribution of the game. *Management Science*, 14(7):486–502.

Harsanyi, J. C. (1975). The tracing procedure: A bayesian approach to defining a solution for n-person noncooperative games. *International Journal of Game Theory*, 4(2):61–94.

Harsanyi, J. C. and Selten, R. (1988). A general theory of equilibrium selection in games. *MIT Press Books*.

Haselton, M. G., Nettle, D., and Murray, D. R. (2015). The evolution of cognitive bias. *The handbook of evolutionary psychology*, pages 968–987.

Hill, K. R., Walker, R. S., Božičević, M., Eder, J., Headland, T., Hewlett, B., Hurtado, A. M., Marlowe, F., Wiessner, P., and Wood, B. (2011). Co-residence patterns in hunter-gatherer societies show unique human social structure. *Science*, 331(6022):1286–1289.

Hochschild, A. R. (2018). *Strangers in their own land: Anger and mourning on the American right*. The New Press.

Hoffman, M., Yoeli, E., and Navarrete, C. D. (2016). Game theory and morality. In *The evolution of morality*, pages 289–316. Springer.

Kalai, E. and Samet, D. (1984). Persistent equilibria in strategic games. *International Journal of Game Theory*, 13(3):129–144.

Kets, W. and Sandroni, A. (2017). A theory of strategic uncertainty and cultural diversity. Technical report, Working paper.

Kohlberg, E. and Mertens, J.-F. (1986). On the strategic stability of equilibria. *Econometrica: Journal of the Econometric Society*, pages 1003–1037.

Kreps, D. M. and Wilson, R. (1982). Sequential equilibria. *Econometrica: Journal of the Econometric Society*, pages 863–894.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, 108(3):480.

Kurzban, R. (2011). *Why everyone (else) is a hypocrite: Evolution and the modular mind*. Princeton University Press.

Kurzban, R. and Christner, J. (2011). Are supernatural beliefs commitment devices for integroup conflict. *The psychology of social conflict and aggression*, 13.

Kurzban, R., Tooby, J., and Cosmides, L. (2001). Can race be erased? coalitional computation and social categorization. *Proceedings of the National Academy of Sciences*, 98(26):15387–15392.

Langer, E. J. and Roth, J. (1975). Heads i win, tails it's chance: The illusion of control as a function of the sequence of outcomes in a purely chance task. *Journal of personality and social psychology*, 32(6):951.

Lecouteux, G. (2018). Bayesian game theorists and non-bayesian players. *The European Journal of the History of Economic Thought*, pages 1420–1454.

Lokhorst, A. M., Werner, C., Staats, H., van Dijk, E., and Gale, J. L. (2013). Commitment and behavior change: A meta-analysis and critical review of commitment-making strategies in environmental research. *Environment and behavior*, 45(1):3–34.

McCullough, M. E., Swartwout, P., Shaver, J. H., Carter, E. C., and Sosis, R. (2016). Christian religious badges instill trust in christian and non-christian perceivers. *Psychology of Religion and Spirituality*, 8(2):149.

Mertens, J.-F. and Zamir, S. (1985). Formulation of bayesian analysis for games with incomplete information. *International Journal of Game Theory*, 14(1):1–29.

Morris, S. (1995). The common prior assumption in economic theory. *Economics & Philosophy*, 11(2):227–253.

Myerson, R. B. (1978). Refinements of the nash equilibrium concept. *International journal of game theory*, 7(2):73–80.

Myerson, R. B. (1991). *Game theory, Analysis of Conflict*. Harvard university press.

Nash, J. (1950). *Non-Cooperative Games*. Doctoral dissertation, Princeton University.

Osborne, M. J. and Rubinstein, A. (1994). *A course in game theory*. MIT press.

Pallak, M. S., Cook, D. A., and Sullivan, J. J. (1980). Commitment and energy conservation. *Applied social psychology annual*.

Pearce, D. G. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica: Journal of the Econometric Society*, pages 1029–1050.

Perea, A. (2014). From classical to epistemic game theory. *International Game Theory Review*, 16(01):1440001.

Pinker, S. (2005). So how does the mind work? *Mind & Language*, 20(1):1–24.

Scott-Phillips, T. C., Dickins, T. E., and West, S. A. (2011). Evolutionary theory and the ultimate–proximate distinction in the human behavioral sciences. *Perspectives on Psychological Science*, 6(1):38–47.

Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4:25–55.

Simler, K. and Hanson, R. (2017). *The elephant in the brain: Hidden motives in everyday life*. Oxford University Press.

Sweeny, K., Melnyk, D., Miller, W., and Shepperd, J. A. (2010). Information avoidance: Who, what, when, and why. *Review of general psychology*, 14(4):340–353.

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., Herrmann, E., Gilby, I. C., Hawkes, K., Sterelny, K., Wyman, E., Tomasello, M., et al. (2012). Two key steps in the evolution of human cooperation: The interdependence hypothesis. *Current anthropology*, 53(6):000–000.

Tooby, J. (2017). Coalitional instincts. https://www.edge.org/response-detail/27168. Accessed: 2019-06-03.

Tooby, J. and Cosmides, L. (2010). Groups in mind: The coalitional roots of war and morality. *Human morality and sociality: Evolutionary and comparative perspectives*, pages 91–234.

Tooby, J., Cosmides, L., and Price, M. E. (2006). Cognitive adaptations for n-person exchange: the evolutionary roots of organizational behavior. *Managerial and Decision Economics*, 27(2-3):103–129.

Trivers, R. (2011). *Deceit and self-deception: Fooling yourself the better to fool others*. Penguin UK.

Von Hippel, W. and Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, 34(1):1–16.