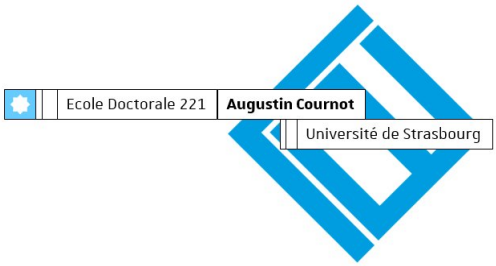
Social incentives and the strategic nature of beliefs and preferences

Arnaud Wolff



UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE AUGUSTIN COURNOT ED 221

BUREAU D'ÉCONOMIE THÉORIQUE ET APPLIQUÉE UMR 7522

THÈSE

pour l'obtention du titre de Docteur en Sciences Économiques

Présentée et soutenue publiquement le 6 Décembre 2022 par

Arnaud Wolff

INCITATIONS SOCIALES ET NATURE STRATÉGIQUE DES CROYANCES ET PRÉFÉRENCES

Préparée sous la direction de Gisèle Umbhauer et Thi Kim Cuong Pham

Composition du jury :

Astrid Hopfensitz	Professeure, EMLyon	Rapporteure
Jean-Baptiste André	Directeur de Recherche CNRS, ENS-PSL	Rapporteur
Eli Spiegelman	Maître de Conférences, Université Bourgogne Franche-Comté	Examinateur
Jocelyn Donze	Professeur, Université de Strasbourg	Examinateur
Gisèle Umbhauer	Maitresse de Conférences HDR, Université de Strasbourg	Directrice de thèse
Thi Kim Cuong Pham	Professeure, Université de Paris-Nanterre	Co-Directrice de thèse

L'Université de Strasbourg n'entend donner aucune approbation, ni improbation aux opinions émises dans cette thèse ; elles doivent être considérées comme propres à leur auteur.

À ma mère et ma grand-mère.

Remerciements

Voici que s'achèvent cinq années de réflexions, doutes, confusions et (rares) illuminations. Cette thèse avait commencé sous la direction de Claire Mainguy et co-dirigée par Thi Kim Cuong Pham, avec pour objectif d'étudier les déterminants de l'action collective dans les villages Indiens. Ayant rapidement remarqué que le projet allait être difficile à mener à terme, je me suis lancé dans la quête d'un nouveau sujet de thèse. Mes idées n'étaient pas très claires, mais je savais que je souhaitais étudier le comportement social en prenant une perspective stratégique. Observant que mes intérêts s'éloignaient des siens, Claire Mainguy a préféré laisser la main à Gisèle Umbhauer, qui est ainsi devenue ma directrice de thèse lors de la seconde année de la thèse. Je veux remercier Claire Mainguy d'avoir accepté d'encadrer mon travail de thèse et de m'avoir laissé par la suite la liberté de poursuivre mes propres intérêts. Je remercie aussi Thi Kim Cuong Pham qui a accepté de co-diriger ma thèse et qui m'a accompagné tout au long de ce travail. Enfin, je souhaite remercier tout particulièrement Gisèle Umbhauer, qui a accepté de prendre la direction de ma thèse alors que celle-ci était déjà bien entamée, qui m'a laissé la liberté de développer mes propres projets, qui a été présente lors de mes nombreux moments de confusion et qui m'a poussé à clarifier mon raisonnement lorsque celui-ci était douteux.

Merci à Jocelyn Donze et Frédéric Koessler qui ont accepté de faire partie de mon comité de suivi de thèse et qui m'ont, pendant trois années, donné de précieux conseils me permettant de peaufiner les arguments présentés dans cette thèse.

Merci à Astrid Hopfensitz, Jean-Baptiste André, Eli Spiegelman et Jocelyn Donze d'avoir accepté d'être membres de mon jury de thèse. C'est un honneur pour moi de pouvoir vous présenter mon travail de thèse et j'anticipe avec enthousiasme vos commentaires et vos suggestions.

Je souhaite aussi remercier tous les membres du Bureau 126 que j'ai eu la chance et le plaisir de rencontrer durant ces cinq dernières années et qui ont significativement amélioré l'expérience du doctorat. Je pense en particulier à Benoît (parti trop tôt), Benoit (toujours à l'écoute du live des Red Hot à Johannesburg en 1991), Huy (mon fidèle partenaire du samedi après-midi), Mali (principale instigatrice de soirées Iraniennes à Strasbourg) et Sila (merci pour les quetsches et mirabelles). Merci à Yann (T.) et Agathe pour tous ces moments de rire et de détresse. Merci aussi à Debora, Enrico, Laetitia, Laté, Quoc et Sam pour tous ces moments d'échange.

Merci à Chouaib, Laurène, Lucas, Thomas et Yann (S.) pour votre soutien intellectuel et émotionnel. Merci à Killian et Andrew de m'avoir accompagné virtuellement durant ces cinq années. Notre conversation "Deadinside" restera à jamais dans

les annales de la NSA.

Merci aux membres de ma famille pour leur présence. Je pense tout particulièrement à ma mère, qui, bien qu'ayant du mal à comprendre comment exactement je contribuais à la société en tapant sur un clavier, gribouillant sur du papier ou lisant un bouquin, n'a pas cessé de me soutenir et a constamment oeuvré afin que je puisse me trouver dans les meilleures conditions possibles. Cette thèse n'aurait pas pu être possible sans toi. Merci aussi à mon grand-père pour ses nombreuses invitations, ses plats raffinés et son inépuisable stock d'histoires à raconter. Toi et Mamie avez toujours été là pour moi et c'est en grande partie vous qui avez façonné la personne que je suis devenue.

Enfin, merci à Kristin (Tintin) d'avoir rendu la seconde partie de la thèse bien plus plaisante que la première. Ta présence quotidienne, ta tendance à trouver les mots justes et ton encouragement sans faille m'ont rassuré, m'ont aidé à réfléchir clairement et m'ont motivé à mener ce projet jusqu'au bout. Stay tuned pour les futures aventures de la compagnie Flausch!

Table of contents

List of Figures

General introduction

The recent decades have observed a surge in findings, from the laboratory and the field, describing how human behavior consistently differs from the predictions of classical economic and game-theoretic models. These findings have led to the emergence of new "behavioral" models, often with a psychological motivation, that aim at better explaining and predicting behavior (Laibson & List 2015). The principal objective of this thesis is to reevaluate this "behavioral" shift in economics and game theory and to propose an alternative grounded in the idea that standard economic and game-theoretic tools can be successfully applied to understand the very deviations from the predictions of classical models, provided that researchers adequately take into consideration the social incentives that individuals face.

Biased Beliefs

The first and main focus of this thesis is what has been termed *biased* beliefs. Standard economic models predict that individuals want their beliefs to be as close as possible to the objective probabilities of outcomes, given that more accurate beliefs lead to better decision-making and greater expected utility (Savage 1954). Additionally, since more accurate beliefs are more desirable, individuals are expected to always welcome and seek new information, given that information about the true state of the world can be used to update beliefs—using Bayes' Rule—and improve decision-making (Stigler 1961). Yet, recent research in psychology and behavioral economics has cast doubt on this purely instrumental perspective on beliefs and information. First, beliefs appear to be biased in *systematic and predictable* ways, with individuals often holding enhanced beliefs about themselves (Taylor & Brown 1988, Kurzban & Aktipis 2007, McKay & Dennett 2009). Second, individuals tend to update their beliefs following good news but fail to do so when receiving bad news—a phenomenon known as *asymmetric updating* (Eil & Rao 2011, Drobner 2022). Third, beliefs in the domains of morality, religion, or politics often appear to be completely unresponsive to evidence, creating *persistent differences* in beliefs among groups of individuals (Kahan 2012, Van Bavel & Pereira 2018). Finally, individuals often actively *avoid* information even when such information can be obtained freely (Dana et al. 2007, Golman et al. 2017). These observations are incompatible with the predictions of standard models and inconsistent with the idea that individuals are uniquely motivated to adopt "objective" beliefs since accuracy-oriented individuals will (given sufficient time) converge to the same beliefs (Geanakoplos & Polemarchakis 1982).

Motivated by these findings, behavioral economists have developed new theoretical frameworks aimed at updating the standard model and accommodating the new evidence (Bénabou 2015, Bénabou & Tirole 2016, Loewenstein & Molnar 2018). The main idea underlying these recent theoretical developments is that individuals care about the beliefs they hold and the information to which they are exposed independently of their contribution to decision-making or their informational value. That is, beliefs (and information) directly enter the individuals' utility function and become objects that we can consume or invest in (Bénabou & Tirole 2016). Deviations from Bayesian rationality—oftentimes called *biased* beliefs—are therefore predicted if such deviations can increase the overall well-being of the individual.

According to the modern *belief-based utility* (BBU) framework (Loewenstein & Molnar 2018), individuals derive pleasure and pain from the beliefs that they hold and these feelings causally influence the adoption or abandonment of specific beliefs. For instance, some news can be undesirable and painful, and individuals might eschew information in order to avoid experiencing pain (Golman et al. 2017). Moreover, benefits stemming from biased beliefs are often modeled as arising from within the individual, such that individuals experience *psychological* benefits from holding biased beliefs. As a matter of example, adopting enhanced beliefs about themselves (e.g., their intelligence, their generosity, their attractiveness, etc.) is assumed to be beneficial for individuals since it improves their well-being (Bénabou & Tirole 2016, Loewenstein & Molnar 2018). Therefore, according to BBU theorists, feelings are conceptualized as *incentives* (or rewards) in and of themselves since they causally influence instrumental behavior (e.g., information avoidance, reality denial, self-signaling, etc.) aimed at experiencing more or less of a specific feeling.

The central part of this thesis will reevaluate some main assumptions of the modern—*belief-based utility*—framework and propose an alternative. The proposed alternative is grounded in the idea that feelings themselves require investigation (Hoffman & Yoeli 2022). Instead of being incentives (or rewards), feelings reflect the workings of emotional mechanisms tracking rewards and punishments in the individual's environment (Rolls 1999, LeDoux 2012). Feelings, therefore, seem to indicate whether what we are doing is good for us or not (Solms 2021). The proposed alternative seeks to understand why a belief might feel good or bad, that is, what are the rewards and punishments (or incentives) that underlie the feeling associated with a specific belief. The focus on incentives—instead of feelings—is meant for the sake of (i) explanation, (ii) prediction, and (iii) intervention. First, taking the feeling itself as an explanation does not clarify why underlying (learning and emotional) mechanisms have placed "positive" or "negative" value on a certain state. Therefore,

concentrating the analysis on rewards and punishments ensures that we provide an ultimate explanation for the behavior (Scott-Phillips et al. 2011). Second, since feelings are functions of rewards and punishments, understanding the incentives that motivate behavior will shed light on the circumstances under which different feelings will emerge. This improved understanding will ultimately allow us to better predict and influence behavior (Rand, Yoeli & Hoffman 2014, Kraft-Todd et al. 2015). Chapter 1 of this thesis will expand on this argument.

In order to illustrate the merits of this approach, two chapters of this thesis aim at describing the incentives underlying (i) the adoption and expression of a given *social identity* and (ii) the nature and form of beliefs about the *self*, respectively. The focus on these categories of beliefs comes from the important role they play for individuals but also for society at large. First, individuals are very attached to their social identity—which defines their membership in social groups—and are usually very eager to display such memberships. Identity signals can help individuals find others with similar beliefs and values, thereby improving coordination and cooperation (Smaldino 2019). Yet, social identity is also thought to be a root cause of various social ills, such as discrimination (Hoff & Pandey 2006) or increased polarization (Iyengar et al. 2019). In economics, two types of models of identity have been developed: *preference-based* and *belief-based* models (Charness & Chen 2020). Preference-based identity models take individual preferences for different social groups as given, while belief-based identity models view identity investments as self-signals. In line with psychological theories, these models tend to emphasize the psychological benefits that individuals derive from their social identity. As described above, my objective in this thesis will be to tie the choice (and expression) of social identity to material and social rewards. This will be done in Chapter 3.

Second, individuals care deeply about their own image, wishing to view themselves as smart, generous, honourable, or attractive people (Bénabou & Tirole 2016, Loewenstein & Molnar 2018). Yet, positively biased perceptions of oneself also lead to overconfidence, which can reduce CEO performance (Malmendier & Tate 2005), make poor leaders (Shipman & Mumford 2011), or even lead nations to war (Johnson 2004). Theoretical work on positively biased beliefs either starts from the assumption that individuals prefer to have a positive self-image (Köszegi 2006, Möbius et al. 2022) or that they use different criteria to evaluate decisions (Van den Steen 2004, Santos-Pinto & Sobel 2005). In Chapter 4 of this thesis, I aim to provide an account of positively biased beliefs which is both grounded in the proposed alternative framework and consistent with the empirical evidence.

Chapters 1, 3, and 4 emphasize the strategic nature of beliefs. In Chapter 5 (co-

authored with Gisèle Umbhauer), we revisit the concept of Sequential Equilibrium by allowing players to have different beliefs at out-of-equilibrium information sets. In line with the ideas developed in Chapter 3, we discuss the notion of distance between the players' beliefs and evaluate different ways to measure that distance. Yet, taking a strategic perspective on beliefs at out-of-equilibrium strategy sets reveals to be intricate. Beliefs can be considered as belonging to the players' strategy set and the beliefs that players announce might be different from the ones they reveal to other players through their actions. These considerations lead us to reconsider the traditional notion of sequential rationality.

Social Preferences

The second focus of this thesis is what has been termed *social preferences*. Researchers have endowed humans with social preferences once it became clear that humans do not only take their own self-interest into account when making decisions in social dilemmas. Instead, humans often cooperate in the Prisoner's Dilemma, they contribute in the Public Goods Game and send money in the Dictator Game. They also reject low offers in the Ultimatum Game, appear averse to inequality and are concerned with fairness. In order to account for the deviations from the predictions of standard models based on individual self-interest maximization, researchers have developed a wide variety of "behavioral" models aimed at better predicting behavior. For instance, a *warm-glow* (Andreoni 1990), or a *preference for altruism* (Levine 1998), have been said to underlie cooperation in social dilemmas.

Nevertheless, these models have been primarily constructed *post-hoc* in order to rationalize observed behavior in the laboratory and the field (DellaVigna 2009). This has led to a proliferation of models aimed at formalizing a particular behavioral observation without an underlying theoretical framework to make sense of it (Fudenberg 2006). Chapter 2, in line with the approach taken in the remainder of the thesis, aims at synthesizing the different "behavioral" models by focusing on the incentives underlying the expression of human social preferences. The detailed outline of the thesis is as follows.

Chapter 1 — Self-Image, Self-Signaling, and the Socially Adapted Mind

According to the modern, *belief-based utility* (BBU) framework (Loewenstein & Molnar 2018), individuals are motivated to improve their *self*-image and they might *self*-signal in order to do so. The objective of this chapter is to reevaluate this

central assumption. The argument is divided into four steps.

First, evolution has bestowed us with action- and outcome-oriented brains whose sole purpose is to improve our prospects of survival and reproduction. Moreover, feelings (typically) can not be considered as either incentives or rewards; rather, feelings reflect the workings of emotional mechanisms which have evolved to deal with opportunities and challenges in our environment effectively. Improving our self-image, or self-signal for its own sake, therefore can not be objectives our minds are striving to achieve: our mind games must necessarily have an effect outside our bodies. Second, what is commonly called the *self* is best seen as the collection of all experiences of life—our thoughts, memories, desires, and sensations—integrated and unified in the mind. That is, the self is not a thing which might differ from what we do, feel or think. Caring about the image of the self must then necessarily imply caring about how we are perceived by others. The desired self-image can, therefore, fruitfully be considered as the desired reputation. Third, while researchers have used subject anonymity in laboratory settings as evidence for *self*-signaling, behavioral spillovers from everyday life to the laboratory are predicted if (i) behavior is controlled by the *model-free* (habitual) system and/or (ii) if codes of conduct have become internalized. In fact, existing empirical evidence supports the idea that individuals bring in the laboratory strategies (heuristics) that have proved useful outside the laboratory. Fourth, more than just behavior, our beliefs, preferences, and intuitions are shaped by learning processes tracking rewards and punishments in our environment. The claim is then the following: what researchers have considered as evidence for self-signaling in the laboratory actually reflects the workings of a psychology well-adapted to the social incentives of everyday life, which spills over when individuals find themselves in unfamiliar environments.

The first chapter of the thesis makes the case that our minds are—by necessity—socially adapted. Rather than being motivated to appear good to themselves, individuals are motivated to appear good to others. Also, individuals do not need to convince themselves but they do need to convince others. When tested in the laboratory, individuals tend to use the same strategies (heuristics) that allow them to maintain a good reputation outside the laboratory. This changing perspective, from the *self* to the social, has significant practical and policy relevance.

Chapter 2 — An Evolutionary Perspective on Social Preferences

Economic theory has taken a "behavioral" turn once it became clear that the traditional models based on individual self-interest maximization could not account

for the fact that individuals take the welfare of others into account when making decisions in the laboratory or the field. That is, standard models were incompatible with the observation that individuals displayed other-regarding (or social) preferences. This has led to the development of a great variety of theoretical models aimed at rationalizing observed behavior in the laboratory or in the field, without any underlying theoretical framework to connect them.

In this chapter, I argue that the theory of social evolution has the necessary *scope* (applying to any type of social interaction) and *power* (clarifying what can and what can not be expected) to provide a useful theoretical framework for human social preferences. Social evolution theory predicts that individuals will be sensitive to the benefits and costs of cooperative acts, and therefore predicts the context-dependent nature of social preferences. Moreover, it illuminates the function of our social emotions, which instead of being fixed traits, are endogenous mechanisms that have evolved to regulate our social relationships with others. Yet, social evolution theory can not predict the content of social behavior in every situation. Therefore, to better understand the wide variation in the expression of social preferences recorded in the laboratory and in the field, we need to take into account the (social) environment in which individuals find themselves.

In fact, a wealth of evidence describes how human social behavior is conditional on the cues present in the local environment, be they socially transmitted or not. I discuss two mechanisms that can help us predict the variable expression of human social preferences: culture/institutions and the local ecology. Overall, this chapter makes the case that to improve our understanding of human social behavior, we need to acknowledge its context-dependent nature and identify the incentives underlying its expression.

Chapter 3 — The Signaling Value of Social Identity in Polarized Environments

Theoretical work on social identity usually emphasizes the psychological benefits (such as self-esteem) that individuals retrieve when adopting and expressing their social identity. Moreover, researchers usually take individual preferences for different social groups as given, leaving open the question of where such preferences come from. In this second chapter, I aim to link the choice of a social identity to (material and social) incentives in an individual's social environment.

More specifically, I argue in this chapter that the choice of a social identity can reveal information about an individual's willingness to cooperate, so that social identity can ultimately signal trustworthiness. The starting point of the argument

is the following: in an environment in which willingness to cooperate is not readily observable, the beliefs and values adopted by individuals are evaluated against the beliefs and values adopted by others whose motives and commitments are commonly known. In this context, by adopting specific beliefs and values, individuals *pool* (or *separate*) from others who adopted similar (or dissimilar) beliefs and values and whose motives and social commitments are known. Adopting a given social identity then essentially signals to others the individual's social commitments and willingness to cooperate. Importantly, the choice of social identity does not depend on individual preferences for different social groups. Rather, it is assumed to be a function of the benefits that the individual is expected to derive from cooperating with different groups of individuals.

The model developed in this chapter makes several predictions which are consistent with empirical evidence. Overall, this chapter makes the case that to better understand the social identity that individuals decide to adopt and express it is important to understand the social incentives that they face (here, cooperation and trustworthiness).

Chapter 4 — The Persuasive Function of Positive Illusions

The stability of positively biased beliefs about oneself—hereafter *positive illusion*—has been explained using three arguments. The first is that positive illusions provide individuals with psychological and health benefits, while the second is that positive illusions result from cognitive or informational biases. These are the arguments typically advanced by psychologists and behavioral economists. The third argument, which is the subject of this chapter, is that individuals adopt positive illusions mainly to persuade others about their (not easily observable) abilities.

The major drawback of the persuasion argument is that it lacks a game-theoretical perspective. In fact, if everyone adopts positive illusions to persuade others about their qualities, then we expect others (receivers) to devalue the signal or to discard it altogether. The objective of this chapter is to investigate whether positive illusions can in fact persuade at equilibrium. To do so, I first develop a two-player "Partner Choice" model in which a Receiver (R) wants to accept a Sender (S) only if S is of higher quality, while S wants to be accepted by R regardless of her type. The question I ask is the following: can S persuade R to accept her by adopting positive illusions (i.e., enhanced beliefs) about her type, even though her type is lower than R's? In a second model, I investigate the stability of positive illusions in a three-player "Community" setting, with one Sender (S) and an audience composed of two Receivers (R_1 and R_2). In this setting, S's objective is to be

seen as better than she is, and both Rs have to decide whether to punish S (or not) after observing the result of a task undertaken by S. Both Rs want to punish S if and only if they expect the other to punish too. The question I ask is the following: can uncertainty about the other R's "punishment threshold" prevent coordinated punishment and allow positive illusions to remain stable at equilibrium?

The results of the theoretical analysis confirm that positive illusions can persuade at equilibrium, even though R is not fooled by S. The existing empirical evidence is in line with the central predictions of both models. As in Chapter 3, the central claim of this chapter is that a better understanding of the social incentives that individuals face (here, persuasion, reputation, and valuable relationships) can help us better understand the beliefs they adopt.

Chapter 5 — Distance in Beliefs and Individually-Consistent Sequential Equilibrium (co-authored with Gisèle Umbhauer)

The concept of Sequential Equilibrium (SE, Kreps & Wilson 1982), often used to solve dynamics games of incomplete information, requires that all players share the same (consistent) beliefs at out-of-equilibrium strategy sets. That is, players need to agree on the numerical value of mathematical artifacts used to generate perturbations of strategy profiles, which are arbitrary by nature.

In this chapter, we extend the concept of Individually-Consistent Sequential Equilibrium (ICSE, Umbhauer & Wolff 2019), which accepts different perturbations systems for different players. Therefore, players can hold different beliefs at out-of-equilibrium strategy sets. We first contrast our solution concept with other commonly used solution concepts, such as Perfect Bayesian Equilibrium (Fudenberg & Tirole 1991) or AGM-consistency (Bonanno 2013, 2016). Since we principally focus on games with $n \geq 3$ players and beliefs are often markers of group membership, a community of players might seek to adopt similar beliefs. This leads us to introduce a notion of distance between beliefs at out-of-equilibrium information sets. This notion of distance can be approached by (i) requiring that players order the perturbed strategies at each out-of-equilibrium information sets in the same way or by (ii) seeking the minimum required changes in payoffs that ensure convergence in beliefs. Yet, analyzing the distance between players' beliefs suggests that beliefs can become objects of choice and therefore belong to the players' strategy set. The beliefs that they announce might be different from the ones that they reveal through their actions, which leads us to discuss the strategic nature of beliefs at out-of-equilibrium strategy sets.

We conclude this chapter by revisiting the notion of sequential rationality in

dynamic games of incomplete information. More than just requiring that players behave optimally at every information set, given their beliefs and the strategies played by other players, we might additionally require that there does not exist another perturbation scheme that is individually-consistent and which provides higher payoffs to the players.

Introduction générale

Ces dernières décennies, nous avons pu assister à une multiplication de travaux de recherche—en laboratoire expérimental et sur le terrain—décrivant comment le comportement humain s'écarte systématiquement des prédictions des modèles classiques en économie et en théorie des jeux. Ces travaux ont conduit à l'émergence de nouveaux modèles "comportementaux", souvent motivés par la psychologie, qui visent à mieux expliquer et prédire le comportement humain (Laibson & List 2015). L'objectif principal de cette thèse est de réévaluer ce virage "comportemental" en économie et en théorie des jeux et de proposer une alternative fondée sur l'idée que les outils classiques d'économie et de théorie des jeux peuvent être appliqués avec succès dans la compréhension des déviations mêmes des prédictions des modèles classiques, à condition que les chercheurs prennent adéquatement en considération les incitations sociales auxquelles les individus font face.

Croyances Biaisées

Le premier et principal sujet de cette thèse concerne ce que l'on appelle les *croyances biaisées*. Les modèles économiques standards prédisent que les individus ont pour objectif que leurs croyances soient aussi proches que possible des probabilités objectives des différents états du monde, étant donné que des croyances plus précises conduisent à une meilleure prise de décision et à une plus grande utilité espérée (Savage 1954). En outre, puisque des croyances plus exactes sont plus souhaitables, on s'attend à ce que les individus acceptent et recherchent de nouvelles informations, étant donné que des informations sur l'état réel du monde peuvent être utilisées pour mettre à jour les croyances—en utilisant la règle de Bayes—et améliorer la prise de décision (Stigler 1961). Pourtant, de récentes recherches en psychologie et en économie comportementale ont jeté le doute sur cette perspective purement instrumentale des croyances et de l'information. Premièrement, certaines croyances semblent être biaisées de manière *systématique et prévisible*, les individus adoptant souvent des croyances positives et optimistes sur leurs traits et caractéristiques (Taylor & Brown 1988, Kurzban & Aktipis 2007, McKay & Dennett 2009). Deuxièmement, les individus ont tendance à mettre à jour leurs croyances après avoir reçu de bonnes nouvelles, mais ne le font pas lorsqu'ils reçoivent de mauvaises nouvelles—un phénomène connu sous le nom de *mise à jour asymétrique* des croyances (Eil & Rao 2011, Drobner 2022). Troisièmement, les croyances dans les domaines de la moralité, de la religion ou de la politique semblent souvent ne pas répondre du tout à l'information, créant ainsi des *différences persistantes* de croyances entre différents

groupes d'individus (Kahan 2012, Van Bavel & Pereira 2018). Enfin, les individus évitent souvent activement l'information, même lorsqu'elle peut être obtenue librement (Dana et al. 2007, Golman et al. 2017). Ces observations sont incompatibles avec les prédictions des modèles standards, et incompatibles avec l'idée que les individus sont uniquement motivés à adopter des croyances "objectives", dans la mesure où des individus strictement motivés par l'exactitude de leurs croyances convergeront nécessairement (avec un temps suffisant) vers les mêmes croyances (Geanakoplos & Polemarchakis 1982).

Motivés par ces résultats, les économistes comportementaux ont développé de nouveaux cadres théoriques visant à mettre à jour le modèle standard et à prendre en compte ces nouvelles découvertes (Bénabou 2015, Bénabou & Tirole 2016, Loewenstein & Molnar 2018). L'idée principale sous-jacente à ces récents développements théoriques est que les individus se soucient des croyances qu'ils entretiennent, et des informations auxquelles ils sont exposés, *indépendamment* de leur contribution à la prise de décision ou de leur valeur informationnelle. En d'autres termes, les croyances (et l'information) entrent directement dans la fonction d'utilité des individus et deviennent des objets que l'on peut consommer ou dans lesquels nous pouvons investir (Bénabou & Tirole 2016). Les déviations par rapport à la rationalité Bayésienne—souvent appelées *croyances biaisées*—sont donc prédites si ces déviations peuvent augmenter le bien-être global de l'individu.

D'après le cadre théorique moderne d'*Utilité Basée sur les Croyances* (UBC), les individus tirent du plaisir et/ou de la douleur des croyances qu'ils entretiennent et ces sentiments influencent de façon causale l'adoption ou l'abandon de certaines croyances. Par exemple, certaines informations peuvent être indésirables et douloureuses, et les individus contournent donc l'information afin d'éviter d'éprouver de la douleur (Golman et al. 2017). De plus, les bénéfices découlant des croyances biaisées sont souvent modélisés comme provenant de l'intérieur de l'individu, de sorte que les individus éprouvent des bénéfices *psychologiques* à avoir des croyances biaisées. Par exemple, l'adoption de croyances positives et optimistes à propos de soi-même (par exemple, son intelligence, sa générosité, son attrait, etc.) est supposée être bénéfique pour les individus car elle améliore leur bien-être (Bénabou & Tirole 2016, Loewenstein & Molnar 2018). Par conséquent, selon les théoriciens de l'UBC, les sentiments sont conceptualisés comme des *incitations* (ou récompenses) puisqu'ils influencent de façon causale le comportement (l'évitement de l'information, le déni de la réalité, l'auto-signal, etc.) visant à ressentir davantage (ou moins) un certain sentiment.

La partie centrale de cette thèse visera à réévaluer certaines hypothèses centrales

du cadre moderne d'*Utilité Basée sur les Croyances* et à proposer une alternative. L'alternative proposée est fondée sur l'idée que les sentiments eux-mêmes nécessitent une explication (Hoffman & Yoeli 2022). Au lieu d'être des incitations (ou des récompenses), les sentiments reflètent le fonctionnement de mécanismes émotionnels qui répondent aux récompenses et punitions dans l'environnement de l'individu (Rolls 1999, LeDoux 2012). Les sentiments semblent donc indiquer si ce que nous faisons est bon pour nous ou non (Solms 2021). L'alternative proposée cherche à déterminer pourquoi une croyance peut être ressentie positivement ou négativement, c'est-à-dire quelles sont les récompenses et les punitions (ou incitations) sous-jacentes au sentiment associé à une certaine croyance. L'accent mis sur les incitations—plutôt que les sentiments—a pour but (i) l'explication, (ii) la prédiction et (iii) l'intervention. Tout d'abord, prendre le sentiment lui-même comme explication ne clarifie pas pourquoi les mécanismes d'apprentissage et émotionnels ont attribué une valeur "positive" ou "négative" à un certain état. Par conséquent, en concentrant l'analyse sur les récompenses et les punitions (les incitations), on s'assure de fournir une explication ultime du comportement (Scott-Phillips et al. 2011). Deuxièmement, puisque les sentiments sont fonctions des récompenses et des punitions, la compréhension des incitations qui motivent le comportement nous éclairera sur les circonstances dans lesquelles différents sentiments émergeront. Cette meilleure compréhension nous permettra en fin de compte de mieux prédire et influencer le comportement (Rand, Yoeli & Hoffman 2014, Kraft-Todd et al. 2015). Le Chapitre 1 de cette thèse développera davantage cet argument.

Afin d'illustrer les mérites de cette approche, deux chapitres de cette thèse visent à décrire les incitations sous-jacentes à (i) l'adoption et l'expression de l'*identité sociale* et (ii) la nature et la forme des croyances sur *soi*, respectivement. L'accent mis sur ces catégories de croyances provient du rôle important qu'elles jouent pour les individus, mais aussi pour la société dans son ensemble. Premièrement, les individus sont très attachés à leur identité sociale—qui définit leur appartenance à certains groupes sociaux—et sont généralement très désireux d'afficher cette appartenance. Les signaux d'identité peuvent aider les individus à trouver d'autres personnes ayant des croyances et des valeurs similaires, améliorant ainsi la coordination et la coopération (Smaldino 2019). Pourtant, l'identité sociale est également considérée comme une cause profonde de divers maux sociaux, tels que la discrimination (Hoff & Pandey 2006) ou la polarisation (Iyengar et al. 2019). En économie, deux types de modèles d'identité sociale ont été développés: les modèles *fondés sur les préférences* et les modèles *fondés sur les croyances* (Charness & Chen 2020). Les modèles d'identité sociale fondés sur les préférences prennent pour ac-

quises les préférences individuelles pour différents groupes sociaux, tandis que les modèles d'identité sociale fondés sur les croyances considèrent les investissements identitaires comme des auto-signaux. Conformément aux théories psychologiques, ces modèles tendent à mettre l'accent sur les bénéfices psychologiques que les individus tirent de leur identité sociale. Comme décrit ci-dessus, mon objectif dans cette thèse sera de lier le choix et l'expression de l'identité sociale à des récompenses (incitations) matérielles et sociales. Cela sera fait dans le Chapitre 3.

Deuxièmement, les individus se soucient beaucoup de leur propre image et souhaitent se voir comme des personnes intelligentes, généreuses, honorables ou attirantes (Bénabou & Tirole 2016, Loewenstein & Molnar 2018). Néanmoins, les croyances sur soi positivement biaisées peuvent conduire à des excès de confiance, qui peuvent réduire les performances des PDGs (Malmendier & Tate 2005), faire de mauvais leaders (Shipman & Mumford 2011) ou même conduire des nations à la guerre (Johnson 2004). Les travaux théoriques sur les croyances positivement biaisées partent soit de l'hypothèse que les individus *préfèrent* avoir une image positive (Köszegi 2006, Möbius et al. 2022), soit qu'ils utilisent différents critères pour évaluer les décisions (Van den Steen 2004, Santos-Pinto & Sobel 2005). Dans le Chapitre 4 de cette thèse, je vise à fournir une nouvelle perspective sur les croyances positivement biaisées qui est à la fois fondée sur le cadre alternatif proposé et cohérente avec les résultats empiriques.

Les chapitres 1, 3 et 4 mettent l'accent sur la nature stratégique des croyances. Dans le Chapitre 5 (co-écrit avec Gisèle Umbhauer), nous revisitons le concept d'équilibre séquentiel en permettant aux joueurs d'avoir des croyances différentes aux ensembles d'information hors équilibre. Dans la lignée des idées développées dans le Chapitre 3, nous discutons l'idée de distance entre les croyances des joueurs et évaluons différentes façons de mesurer cette distance. Néanmoins, adopter une perspective stratégique sur les croyances aux ensembles d'information hors équilibre s'avère complexe. Les croyances peuvent être considérées comme appartenant à l'ensemble de stratégies des joueurs et les croyances que les joueurs annoncent peuvent être différentes de celles qu'ils révèlent aux autres joueurs à travers leurs actions. Ces considérations nous amènent à reconsidérer la notion traditionnelle de rationalité séquentielle.

Préférences Sociales

Le deuxième axe de cette thèse concerne les *préférences sociales*. Les chercheurs ont doté les humains de préférences sociales lorsqu'il est apparu que les humains ne tiennent pas uniquement compte de leur propre intérêt lorsqu'ils prennent des

décisions dans des dilemmes sociaux. Au contraire, les humains coopèrent souvent dans le Dilemme du Prisonnier, ils contribuent dans le jeu du Bien Public et ils envoient de l'argent dans le jeu du Dictateur. Ils rejettent également les offres basses dans le jeu de l'Ultimatum, semblent avoir une aversion pour l'inégalité et sont préoccupés par l'équité. Afin d'expliquer ces écarts par rapport aux prédictions des modèles standard basés sur la maximisation de l'intérêt individuel, les chercheurs ont développé une grande variété de modèles "comportementaux" visant à mieux prédire le comportement social humain. Par exemple, le *warm-glow* (Andreoni 1990), ou une *préférence pour l'altruisme* (Levine 1998), seraient à la base de la coopération dans les dilemmes sociaux.

Néanmoins, ces modèles ont été principalement construits *post-hoc* afin de rationaliser le comportement observé en laboratoire expérimental et sur le terrain (DellaVigna 2009). Cela a conduit à une prolifération de modèles visant à formaliser une particulière observation comportementale, sans cadre théorique sous-jacent pour lui donner un sens (Fudenberg 2006). Le Chapitre 2 de cette thèse, en accord avec l'approche adoptée dans le reste de la thèse, vise à synthétiser les différents modèles "comportementaux" en se focalisant sur les incitations sous-jacentes à l'expression des préférences sociales humaines. Le plan détaillé de la thèse est le suivant.

Chapitre 1 — L'image de soi, l'auto-signal et l'esprit socialement adapté

Selon le cadre moderne d'*Utilité Basée sur les Croyances* (UBC), les individus sont motivés à améliorer l'image qu'ils ont d'eux-mêmes, et ils peuvent émettre des signaux envers eux-mêmes (des auto-signaux) pour y parvenir. L'objectif de ce chapitre est de réévaluer cette hypothèse centrale. L'argumentation se divise en quatre étapes.

Premièrement, le processus d'évolution par sélection naturelle nous a doté de cerveaux orientés vers l'action dont le seul but est d'améliorer nos chances de survie et de reproduction. De plus, les sentiments ne peuvent en général pas être considérés comme des incitations ou des récompenses; ils reflètent plutôt le fonctionnement de mécanismes émotionnels qui ont évolué afin de nous permettre de répondre efficacement aux opportunités et aux défis de notre environnement. L'amélioration de l'image de soi, ou l'auto-signal, ne peuvent donc pas être des objectifs que notre esprit s'efforce d'atteindre: les jeux auxquels nous jouons (avec nous-mêmes) doivent nécessairement avoir un effet en dehors de notre corps. Deuxièmement, ce que l'on appelle communément le *soi* est mieux vu comme la collection de toutes les expériences de la vie—nos pensées, nos souvenirs, nos désirs et nos sensations—intégrées

et unifiées dans l'esprit. Autrement dit, le soi n'est pas une chose qui pourrait être différente de ce que nous faisons, ressentons ou pensons. Se soucier de l'image de soi implique donc nécessairement de se soucier de la façon dont nous sommes perçus par les autres. L'image de soi souhaitée peut donc être considérée comme la réputation (image sociale) souhaitée. Troisièmement, alors que les chercheurs ont utilisé l'anonymat des sujets en laboratoire comme preuve de l'existence de l'auto-signal, les "transferts" comportementaux de la vie quotidienne vers le laboratoire sont prédits si (i) le comportement est contrôlé par le système habituel et/ou (ii) si les codes de conduite ont été internalisés. Il s'avère que les preuves empiriques existantes soutiennent l'idée que les individus apportent dans le laboratoire des stratégies (heuristiques) qui se sont avérées utiles en dehors du laboratoire. Quatrièmement, plus que le comportement, nos croyances, préférences et intuitions sont façonnées par des processus d'apprentissage qui répondent aux récompenses et punitions dans notre environnement. L'argument est alors le suivant: ce que les chercheurs ont considéré comme preuve d'auto-signal dans le laboratoire reflète le fonctionnement d'une psychologie bien adaptée aux incitations sociales de la vie quotidienne, qui est "transférée" lorsque les individus se retrouvent dans des environnements non familiers tels que le laboratoire.

Le premier chapitre de la thèse démontre que nos esprits sont—par nécessité—socialement adaptés. Plutôt que d'être motivés à paraître bons envers eux-mêmes, les individus sont motivés à paraître bons aux yeux des autres. De même, les individus n'ont pas besoin de se convaincre eux-mêmes, mais ils ont besoin de convaincre les autres. Lorsqu'ils sont testés en laboratoire, les individus ont tendance à utiliser les mêmes stratégies (heuristiques) qui leur permettent de conserver une bonne réputation en dehors du laboratoire. Ce changement de perspective, du *soi* vers le social, a des conséquences pratiques et politiques importantes.

Chapitre 2 — Une perspective évolutionnaire sur les préférences sociales

La théorie économique a pris un tournant "comportemental" lorsqu'il est apparu que les modèles traditionnels, basés sur la maximisation de l'intérêt individuel, ne pouvaient pas rendre compte du fait que les individus prennent en compte le bien-être des autres lorsqu'ils prennent des décisions dans des dilemmes sociaux. En d'autres termes, les modèles standards étaient incompatibles avec l'observation que les individus affichaient des préférences sociales. Cela a conduit au développement d'une grande variété de modèles théoriques visant à rationaliser les comportements observés en laboratoire ou sur le terrain, sans qu'aucun cadre théorique sous-jacent

ne les relie.

Dans ce chapitre, je soutiens que la théorie de l'évolution sociale possède la *portée* (s'appliquant à tout type d'interaction sociale) et la *capacité prédictive* (clarifiant ce qui peut et ce qui ne peut pas être attendu) nécessaires pour fournir un cadre théorique utile aux préférences sociales humaines. La théorie de l'évolution sociale prédit que les individus seront sensibles aux bénéfices et aux coûts des actes de coopération, et prédit donc la nature dépendante au contexte des préférences sociales. En outre, elle éclaire la fonction de nos émotions sociales qui, au lieu d'être des traits fixes, sont des mécanismes endogènes qui ont évolué afin de réguler nos interactions sociales avec les autres. Néanmoins, la théorie de l'évolution sociale ne peut pas prédire le contenu du comportement social dans chaque situation. Par conséquent, pour mieux comprendre les larges variations dans l'expression des préférences sociales enregistrées dans le laboratoire et sur le terrain, nous devons tenir compte de l'environnement (social) dans lequel se trouvent les individus.

De nombreux travaux décrivent comment le comportement social humain est dépendant d'informations présentes dans l'environnement local, qu'elles soient socialement transmises ou non. Je discute deux mécanismes qui peuvent nous aider à prédire l'expression variable des préférences sociales: la culture/les institutions et l'écologie locale. Dans l'ensemble, ce chapitre montre que pour améliorer notre compréhension du comportement social humain, nous devons reconnaître sa nature dépendante au contexte et identifier les incitations sous-jacentes à son expression.

Chapitre 3 — La valeur de signal de l'identité sociale dans les environnements polarisés

Les travaux théoriques sur l'identité sociale mettent généralement l'accent sur les bénéfices psychologiques (tels que l'estime de soi) que les individus retirent en adoptant et en exprimant leur identité sociale. En outre, les chercheurs considèrent généralement les préférences des individus pour différents groupes sociaux comme données, laissant ouverte la question de l'origine de ces préférences. Dans ce deuxième chapitre, je cherche à lier le choix d'une identité sociale aux incitations (matérielles et sociales) dans l'environnement social d'un individu.

Plus précisément, je soutiens dans ce chapitre que le choix d'une identité sociale peut révéler des informations sur la volonté de coopérer d'un individu, de sorte que l'identité sociale peut être un signal de fiabilité. Le point de départ de l'argument est le suivant: dans un environnement où la volonté de coopérer n'est pas facilement observable, les croyances et les valeurs adoptées par les individus sont évaluées par rapport aux croyances et aux valeurs adoptées par d'autres personnes

dont les valeurs et les motivations sont communément connues. Dans ce contexte, en adoptant des croyances et des valeurs spécifiques, les individus s'*associent* (respectivement se *séparent*) avec d'autres ayant adopté des croyances et des valeurs similaires (respectivement différentes) et dont les valeurs et les motivations sociales sont connues. L'adoption d'une certaine identité sociale signale alors essentiellement aux autres les engagements sociaux de l'individu et sa volonté de coopérer. Il est important de noter que le choix de l'identité sociale ne dépend pas des préférences de l'individu pour différents groupes sociaux. Ce choix est plutôt supposé être fonction des bénéfices que l'individu peut tirer de la coopération avec différents groupes d'individus.

Le modèle développé dans ce chapitre fait plusieurs prédictions qui sont cohérentes avec les travaux empiriques existant. Globalement, ce chapitre montre que pour mieux comprendre l'identité sociale que les individus décident d'adopter et d'exprimer, il est important de comprendre les incitations sociales auxquelles ils sont confrontés (ici, la coopération et la confiance).

Chapitre 4 — La fonction persuasive des illusions positives

La stabilité des croyances sur soi-même positivement biaisées—aussi appelées *illusions positives*—a été expliquée à l'aide de trois arguments. Le premier argument est que les illusions positives procurent aux individus des bénéfices psychologiques et/ou en termes de santé, alors que le second argument est que les illusions positives résultent de biais cognitifs ou informationnels. Ce sont les arguments généralement avancés par les psychologues et les économistes comportementaux. Le troisième argument, qui fait l'objet de ce chapitre, est que les individus adoptent des illusions positives principalement pour persuader les autres de leurs qualités (difficilement observables).

Le principal inconvénient de l'argument de la persuasion est qu'il ne s'inscrit pas dans la logique de la théorie des jeux. En effet, si chacun adopte des illusions positives pour persuader les autres de leurs qualités, nous nous attendons à ce que les autres (les récepteurs) dévaluent le signal ou le rejettent complètement. L'objectif de ce chapitre est d'étudier si les illusions positives peuvent effectivement persuader à l'équilibre. Pour ce faire, je développe d'abord un modèle de "choix du partenaire" à deux joueurs dans lequel un récepteur (R) veut accepter un expéditeur (S) uniquement si S est de meilleure qualité que lui, tandis que S veut être accepté par R quel que soit son type. La question que je pose est la suivante: S peut-il persuader R de l'accepter en adoptant des illusions positives (c'est-à-dire des croyances positivement biaisées) sur son type, même si son type est inférieur à celui de R ? Dans un second

modèle, j'étudie la stabilité des illusions positives dans un cadre de "communauté" à trois joueurs, avec un émetteur (S) et une audience composée de deux récepteurs (R_1 et R_2). Dans ce contexte, l'objectif de S est d'être perçu comme meilleur qu'il ne l'est, et les deux R doivent décider de punir S (ou non) après avoir observé le résultat d'une tâche entreprise par S. Les deux R veulent punir S si et seulement s'ils s'attendent à ce que l'autre punisse aussi. La question que je pose est la suivante: l'incertitude sur le "seuil de punition" de l'autre R peut-elle empêcher la punition coordonnée et permettre aux illusions positives de rester stables à l'équilibre ?

Les résultats de l'analyse théorique confirment que les illusions positives peuvent persuader à l'équilibre, même si R n'est pas dupé par S. Les travaux empiriques existant sont en accord avec les prédictions centrales des deux modèles. Comme dans le Chapitre 3, l'argument principal de ce chapitre est qu'une meilleure compréhension des incitations sociales auxquelles les individus sont confrontés (ici, la persuasion, la réputation et la coopération) peut nous aider à mieux comprendre les croyances qu'ils adoptent.

Chapitre 5 — Distance entre les croyances et équilibre séquentiel individuellement-cohérent

Le concept d'équilibre séquentiel (ES, Kreps & Wilson 1982), souvent utilisé pour résoudre les jeux dynamiques à information incomplète, exige que tous les joueurs partagent les mêmes croyances (cohérentes) aux ensembles d'information hors équilibre. En d'autres termes, les joueurs doivent se mettre d'accord sur la valeur numérique d'artefacts mathématiques utilisés pour générer des perturbations des profils stratégiques, qui sont par nature arbitraires.

Dans ce chapitre, nous étendons le concept d'équilibre séquentiel individuellement-cohérent (ESIC, Umbhauer & Wolff 2019), qui accepte différents systèmes de perturbation pour différents joueurs. Par conséquent, les joueurs peuvent avoir des croyances différentes aux ensembles d'information hors équilibre. Nous commençons par opposer notre concept de solution à d'autres concepts de solution couramment utilisés, tels que l'équilibre Bayésien parfait (Fudenberg & Tirole 1991) ou la cohérence AGM (Bonanno 2013, 2016). Puisque nous nous concentrons principalement sur les jeux avec $n \geq 3$ joueurs et puisque les croyances sont souvent des marqueurs d'appartenance à un groupe, une communauté de joueurs pourrait chercher à adopter des croyances similaires. Ceci nous amène à introduire une notion de distance entre les croyances aux ensembles d'information hors équilibre. Cette notion de distance peut être approchée (i) en exigeant que les joueurs ordonnent les stratégies perturbées à chaque ensemble d'information hors équilibre de la même

manière ou (ii) en recherchant les changements minimaux requis dans les gains qui assurent la convergence des croyances. L'analyse de la distance entre les croyances des joueurs suggère que les croyances peuvent devenir des objets de choix et donc appartenir à l'ensemble de stratégies des joueurs. Les croyances qu'ils annoncent peuvent être différentes de celles qu'ils révèlent par leurs actions, ce qui nous amène à discuter la nature stratégique des croyances aux ensembles d'information hors équilibre.

Nous concluons ce chapitre en revisitant la notion de rationalité séquentielle dans les jeux sous forme extensive. En plus d'exiger que les joueurs se comportent de manière optimale à chaque ensemble d'information étant donné leurs croyances et les stratégies jouées par les autres joueurs, nous pourrions également exiger qu'il n'existe pas d'autre système de perturbations qui soit individuellement-cohérent et qui fournisse des gains plus élevés aux joueurs.

Chapter 1

Self-Image, Self-Signaling, and the Socially Adapted Mind

Summary

Recent work in behavioral economics has suggested that individuals derive utility from the beliefs that they hold. More specifically, individuals are assumed to be motivated to improve their *self-image*. In order to maintain a positive self-image, their behavior needs to be consistent with their beliefs about themselves, which leads to *self-signaling*. The objective of this chapter is to reevaluate the idea that (i) individuals care about their self-image and (ii) individuals self-signal. I first argue that the desired self-image is best seen as the desired reputation. Then, I defend the idea that what appears to be self-signaling in the laboratory reflects the workings of a psychology well-adapted to the social incentives of everyday life, which spills over when individuals find themselves in new, contrived environments.

Classification

JEL Classification: C70, D01, D91

Keywords: Beliefs, Self-Image, Self-Signaling, Social Image, Social Incentives

1.1 Introduction

A growing literature in behavioral economics is advancing the idea that individuals derive utility directly from the beliefs that they hold (Bénabou & Tirole 2016, Loewenstein & Molnar 2018). This research program aims at explaining the systematic deviations from the predictions of standard economic models that we observe in the laboratory or in the field (Molnar & Loewenstein 2021). For instance, the standard (*Subjective Expected Utility*—SEU) model predicts that individuals will want their beliefs to be as accurate as possible since more accurate beliefs lead to better decision-making. Another prediction of the standard model is that individuals will always seek and welcome new information and rationally update their beliefs in accordance with the evidence since this is the best way to maximize expected utility. Yet, in the laboratory and in the field, we tend to observe systematic and persistent deviations from epistemic rationality, with individuals often updating their beliefs in response to good but not to bad news, a phenomenon known as *asymmetric updating* (Eil & Rao 2011). Moreover, individuals often avoid decision-relevant information, especially when they expect the information to be painful, which appears to lead to sub-optimal decision-making (Golman et al. 2017). According to the *belief-based utility* (BBU) framework (Loewenstein & Molnar 2018), these systematic deviations can be explained by the fact that individuals derive utility from what they believe, which is something the SEU model does not take into account. Since beliefs enter the utility function, the BBU framework predicts deviations from epistemic rationality if such deviations can increase the overall welfare of the individual.

Central to the BBU framework is the idea that individuals care about their *self-image* (Loewenstein & Molnar 2018), which implies that they derive ego-utility (or pleasure) from thinking about themselves as competent, generous, honourable or moral persons. This is in line with work from social psychology, according to which individuals adopt positive views about themselves in order to improve their mental health, promote their well-being and protect their self-esteem (Taylor & Brown 1988). Yet, according to BBU theorists, the self is not easily fooled: individuals can not just decide to adopt positive beliefs about themselves, but have to behave in such a way as to convince themselves that they are competent, generous, honourable, or moral persons. Therefore, individuals are essentially assumed to play games with themselves and to behave in such a way as to maintain and enhance their self-image (Bénabou & Tirole 2011, Bodner & Prelec 2003, Grossman & Van der Weele 2017). Such *self-signaling* is assumed to underlie a wide range of empirical observations from the laboratory and the field (Bénabou 2015).

The objective of this chapter is to reevaluate the idea that (i) individuals care about their self-image and (ii) individuals self-signal. This will be done in several steps. The first step involves an evolutionary argument against the idea that self-image, or self-signaling, might be ends in themselves. Evolution has endowed us with action- and outcome-oriented brains that ultimately "care" about survival and reproductive outcomes. Feeling good about oneself, or trying to convince oneself about something for its own sake, can not be objectives our minds are striving to achieve. Moreover, emotions are functional states which have evolved to deal with recurring environmental challenges that require an effective, adaptive response (Adolphs & Anderson 2018). The reason why humans consciously experience (*feel*) emotions is still a matter of debate (LeDoux 2012), but it is clear that emotional mechanisms are closely tracking rewards and punishments (and changes therein) in the individual's environment (Rolls 1999). Therefore, emotional mechanisms are useful only to the extent that they motivate us to behave in adaptive ways, which requires that we *act* on the world, and not play games with(in) ourselves.

The second step involves a discussion of the meaning of *self*. The self, while not an illusion, is best seen as a trick played by our minds, which creates unity out of our experiences, desires, beliefs, and sensations (Baggini 2011, Seth 2021). That is, rather than being a thing or an entity which *has* all the experiences of life, the self actually *is* the collection of all theses experiences. The self is therefore a process, constantly evolving (Baggini 2011, Baumeister 2022). This suggests that there is in fact no entity which represents the "True" self (Baumeister 2019). Rather, the "True" self appears to be an ideal, a guiding idea towards which we strive and which mainly responds to what society values (Baumeister 2022). The *self-image* can therefore fruitfully be seen as the *social image*, or desired reputation. This shift in emphasis, from inside the individual (self) to outside (social), is consistent with the idea that our brains are action- and outcome-oriented, so that individuals are ultimately motivated to improve their reputation and to be valued by others.

The third step involves a discussion of the learning processes underlying behavior. More specifically, the discussion will revolve around the *reinforcement learning* (RL) and *social learning* (SL) frameworks. Both the RL and the SL frameworks predict spillovers from everyday life to the laboratory if (i) behavior is controlled by the *model-free* (habitual) system and/or (ii) if codes of conduct have become internalized. The fourth and last step describes how, more than just behavior, learning processes tracking rewards and punishments also shape our beliefs, preferences, and intuitions (Hoffman & Yoeli 2022). The existing empirical evidence is in line with the idea that behavior in the laboratory reflects patterns of behavior in everyday

life, such that individuals bring in the laboratory strategies (heuristics) that have proved useful outside the laboratory.

The final part of the chapter reviews the self-signaling framework in light of the arguments exposed in the chapter. Self-signaling interpretations require that we accept that (i) individuals at times know their true preferences, but at other times can not sufficiently introspect, that (ii) they (the *decision-maker* self) can manufacture "diagnostic" signals which they (the *observer* self) then interpret as impartial, and that (iii) a positive self-image is a fundamental motive which warrants such mind games. I argue that the claim that laboratory experiments tap into a psychology well-adapted to the social incentives of everyday life is more parsimonious. For instance, the lack of an audience in laboratory experiments has been taken as evidence that ignorance of information about the social impact of one's action serves to obfuscate (in the eyes of the observer) the choice the individual (decision-maker) would have made, had they received the information (Grossman & Van der Weele 2017). Yet, if willful avoidance of information allows individuals to maintain plausible deniability outside the laboratory and individuals learn this through reinforcement or observation, then we can expect individuals to adopt the same strategy in the laboratory, even though they are by themselves.

The chapter concludes by discussing the practical and policy relevance of this changing perspective, from the self to the social. Rather than appeals to self-image, the development of institutions and organizations incentivizing cooperation is required in order to promote individually-costly prosocial behavior and tackle society's most pressing issues.

1.2 Evolution only cares about outcomes

Social psychologists have long defended the idea that individuals adopt positive beliefs about themselves in order to improve their psychological well-being (Taylor & Brown 1988). Recent work in behavioral economics is following that insight. According to the *belief-based utility* (BBU) framework, individuals derive utility directly from the beliefs that they hold, such that "people will have an incentive to hold beliefs that make them feel good" (Loewenstein & Molnar 2018, p.167). Central to this framework is the idea that individuals care about how they see themselves (Bénabou & Tirole 2016, Loewenstein & Molnar 2018, Molnar & Loewenstein 2021). For instance, Bénabou & Tirole (2016, p.146) write that "seeing oneself as smart, attractive, and good is intrinsically more satisfying than the reverse", while Loewenstein & Molnar (2018, p.166) suggest that "what people really care about is their

self-image: they want to view themselves as generous, honourable people". The main idea is that feelings become *incentives* (or rewards) in and of themselves, in the sense that they causally influence instrumental behavior specifically aimed at experiencing more or less of the specific feeling.

The objective of this section is twofold. First, I want to argue against the idea that improving one's self-image, or self-signal, can be ends in themselves. Second, I want to argue against the idea that feelings (typically) are incentives.

1.2.1 Brains are action-oriented

The idea that individuals ultimately care about their self-image, or might signal to themselves, seems intuitively plausible but does not appear to stand upon further scrutiny. Our brains have evolved in order to "(1) allow us to identify things in the world; (2) tell us what attitudes and goals to have with respect to them; and (3) move our bodies about in ways that are appropriate to those goals" (Barrett 2014, p.17). In fact, brains are fundamentally *action-oriented*: the principal role of the nervous system is to allow organisms to remain within a range of desirable states, notably by controlling behavior (Cisek 2019). It follows that to evolve in the population, brain mechanisms must necessarily cause (i) a better regulation of the organism's internal milieu (through metabolism) and/or (ii) a greater probability of survival or reproduction (Sterling & Laughlin 2015). This suggests that improving one's self-image to feel better about oneself, or self-signal to convince oneself of something, can not be objectives our minds are striving to achieve since neither contributes to a better regulation of the internal milieu or a greater probability of survival or reproduction.[1] For such mental adaptations to evolve, they would need to have an effect outside the body, onto the world, since how the mind feels about itself is invisible to the process of evolution by natural selection.

In any case, designing an organism capable of choosing what to believe in order to feel better about itself is an evolutionary dead-end. Marvin Minsky (cited in Kurzban 2012) accurately summarizes this point when he writes:

> If we could deliberately seize control of our pleasure systems, we could reproduce the pleasure of success without the need for any actual accomplishment. And that would be the end of everything. (Minsky 1988, p.68)

[1] While some researchers have argued that positively biased beliefs can improve health outcomes and therefore contribute to a better regulation of the internal milieu (McKay & Dennett 2009), the empirical literature has failed to find such positive outcomes (Coyne & Tennen 2010, Sedikides 2022).

What this argument suggests is that the feeling of wanting to improve one's self-image, or the feeling of wanting to convince oneself of something, must necessarily have an influence on how we behave, on how we act on the world. *We* (our *selves*) can not be the ultimate targets of our mind games.

1.2.2 Feelings are (typically) not incentives

The argument in this section requires that we differentiate *feelings* from the underlying *emotions* (or *survival circuits*) that contribute to their emergence. Emotions are highly conserved across species (Adolphs & Anderson 2018), and include at minimum circuits involved in defense, maintenance of energy and nutritional supplies, fluid balance, thermoregulation, and reproduction (LeDoux 2012). For instance, the defense circuit evolved in order to detect threats in the environment and coordinate behavioral and physiological responses. In humans, social emotions such as shame, guilt, pride, gratitude or compassion are similarly assumed to solve adaptive problems related to cooperative social living (Beltran et al. 2022). These circuits are tuned to and activated by "emotional" stimuli, which are either innate (e.g., the innate aversion to spiders or looming shadows) or learned (e.g., the value placed on social partners). An important aspect of emotional states is that they are valenced, namely, that they specify whether the stimuli are to be approached or avoided (Rolls 1999). "Emotional" stimuli can, therefore, usefully be defined as *incentives*, which motivate (approach or withdrawal) instrumental behavior (LeDoux 2012).

Feelings are defined as the conscious experience of an emotion. The reason why humans consciously *feel* emotions is still a matter of debate, and this debate is deeply tied to the question of the origins of consciousness (Adolphs & Anderson 2018, Chapter 10). What matters for the purposes of this discussion is that emotions have essentially evolved in order to solve adaptive problems. In fact, leading theorists have suggested that (i) feelings are cognitively constructed, when body feedback is integrated with environmental stimuli, memories and expectations (Barrett 2017, LeDoux & Brown 2017), that (ii) subjective experiences constitute a "*post hoc* 'commentary' on the sensory representation itself" (Everitt & Robbins 2005, p.1483), or that (iii) feelings "are the subjective aspects of predictions about the causes of interoceptive signals" (Seth 2021, p.209). According to this view, what we experience as feelings either reflects, or constitutes a prediction of the body's physiological condition.

Functional accounts of the role of feelings, on the other hand, suggest that (i) hedonic experiences assign values to stimuli on the basis of the experienced affect, with the nature of affect (or feelings) determined by the workings of a mechanistic

Stimulus-Response (S/R) psychology which places "positive" and "negative" value to biologically important reinforcers (Dickinson & Balleine 2010),[2] or that (ii) feelings signal needs (prediction errors) which subsequently motivate voluntary behavior directed at resolving the need (Solms 2021).[3] This work suggests that instead of being incentives (or rewards), feelings *reflect* the underlying workings of emotional mechanisms which track rewards and punishments in the individual's environment. In this view, the feelings that we experience are neither things towards which we work nor rewards in themselves, but rather signals about (i) our current condition and/or (ii) the need to "do work"—that is, maintain or improve our current state. Therefore, if something (e.g., a belief) feels good, a thorough explanation requires that we specify why (in terms of tangible benefits and costs for the individual) that something feels good. Taking the feeling itself as an explanation for the behavior leaves unanswered the question of why underlying emotional mechanisms have placed "positive" or "negative" value on such a state.

This observation, of course, does not imply that individuals never behave in such a way as to feel a certain way. For instance, individuals watch horror movies to feel afraid, listen to sad songs to feel nostalgic, watch magicians to feel surprised, or delay rewards to experience the feeling of anticipation (Loewenstein 1987). In such instances, feelings are incentives *and* rewards, given that the prospect of experiencing the feeling itself drives behavior, and the behavior is maintained solely due to the ensuing feeling experience. Yet, there is a case to be made that this class of behavior is not as fitness-relevant as, for instance, one's belief—and subsequent behavior—about one's own health, one's mating value, or one's skills and competence. People typically know that they are safe in a movie theater, that the magician's deception will not hurt them, or that delaying a small reward will not make a huge difference. However, when it comes to one's health, mating prospects or skills, it would be sub-optimal, as described in Section 1.2.1, to let the sole prospect of experiencing a given feeling drive behavior.

1.3 What is the *self*?

The *self* features prominently in everyday discussions as well as in academic debates. Concepts such as *self*-awareness, *self*-esteem, *self*-regulation, *self*-actualization or

[2] This is in line with the fact that reinforcement happens at the molecular and cellular levels, and that feelings are not necessary for reinforcement to occur (LeDoux 2015, Chapter 4).

[3] According to Solms (2021), since feelings signal need, the ideal state is actually the state in which organisms do *feel nothing*, since this is a state of certainty where all needs are met automatically (unconsciously).

self-presentation are routinely used in scientific papers (Leary 2004). But what exactly does the self represent? Defining the self has been (and remains) a daunting task for psychologists and philosophers, but a common view is that the self is a "thing", a stable core, which *has* all the experiences of life (Baggini 2011). This view is reflected in the common idea of a "True" self, a supposed entity which represents who individuals really are and towards which individuals strive.

The objective of this section is twofold. First, I want to argue against the idea that the self is a "thing". Second, I want to propose that *self*-image is best viewed as *social* image, or desired reputation.

1.3.1 The *self* is not a thing

Individuals typically experience life in the first-person view with a strong sense of psychological unity and continuity. This creates the intuition that there must exist something which has all these experiences. That something is typically thought to be *us*, *ourselves*. This strong intuition gave birth to the idea of the *self*, an entity, a core, or an essence which experiences life. In this view, the self has thoughts, the self has desires, and the self has worries; it is a stable entity that remains relatively unchanged throughout life. This intuition underlies the need to self-actualize (realize our true potential) or the need to be authentic, true to ourselves. It also underlies the idea that there exists a "True" self, an entity which *really* is us.

But if the self is a "thing" or an entity, then where does it reside? In *The Ego Trick*, Baggini (2011) persuasively argues against the idea that the self can be found in our body, in our brain, in our memories or in our dispositions. There is no single part that contains our inner self. For instance, there is no center in the brain where "it all comes together", where the self might reside. Rather, our bodies, brains, memories and dispositions all *contribute* to the self. In fact, according to Baggini (2011), the unity of the self is a trick played by our minds, which he calls the *Ego Trick*:

> The trick is to create something which has a strong sense of unity and singleness from what is actually a messy, fragmented sequence of experiences and memories, in a brain which has no control centre. (Baggini 2011, p.119)

Similarly, in *Being You*, Seth (2021) writes that:

> The self is not an immutable entity that lurks behind the windows of the eyes, looking out into the world and controlling the body as a pilot controls a plane. The experience of being me, or of being you, is

> a perception itself—or better, a collection of perceptions. (Seth 2021, p.181)

Therefore, instead of being a thing or an entity which underlies all the experiences of life, the self actually *is* the collection of all these experiences. There is no unified core; the self is a bundle-like system, a collection of thoughts, memories and desires which are integrated in the mind and which constantly evolves. It is therefore a feature of selves that they evolve throughout life as new memories and experiences are created and yet that we still feel that we are the same person over time. This view helps explain why damage to one part (e.g., to the body following an accident, to the brain following brain damage, or to memories following a neurodegenerative disease) does not completely destroy the self, except in the most extreme cases.

This view is also shared by Roy Baumeister, a leading researcher on the self. In *The Self Explained*, he writes:

> [T]he self is something the brain does rather than something that is (exists) inside it. It's a process rather than a thing, not unlike life itself. (Baumeister 2022, p.45)

The view of the self as a process challenges the idea that there exists some entity which remains relatively stable throughout life, which *really is us*, and which might be different from how we think or behave. Summarizing the large literature on authenticity, Baumeister (2019, p.143) concludes that "[t]he idea of a true self different from one's actual actions, roles, and experiences is probably indefensible". But if there is no core, no entity representing the self, then what does the *self*-image represent?

1.3.2 *Self*-image as *social* image

If the self is the collection of actions, desires, thoughts, memories and experiences—all integrated in the mind—then being self-aware implies that thoughts are directed towards the self. One's *self*-image therefore represents how one perceives one's actions, desires, memories and experiences. Individuals seem to greatly value their self-image, often trying to enhance their perceptions of themselves. Yet, as described in Section 1.2.1, caring about how one perceives oneself can not be an end in itself; the energy spent caring about and trying to improve one's self-image must necessarily lead to outcomes in the world (e.g., a change in behavior). Moreover, since there is no point to perception without action, our perception of ourselves must necessarily be in the service of control and regulation of behavior (Seth 2021).

This suggests that caring about one's self-image ultimately implies caring about one's *social* image (or *reputation*), since one's image in the eyes of others actually has important consequences on how they value and treat us. Improving our social image is a plausible objective our minds would be designed to achieve and would justify our constant worry about how we perceive ourselves. In this view, the need to improve one's self-image is driven by the need to improve one's image in the eyes of others, with the desired self-image representing the desired reputation (Baumeister 2019, 2022). In line with this idea, some research has shown that individuals feel most authentic (true to themselves) when they behave in accordance with what society values (Fleeson & Wilt 2010, Sheldon et al. 1997).[4]

This shift in emphasis, from inside the individual (the self) to outside the individual (the social) is therefore in line with the argument exposed in Section 1.2 and in line with an influential theory called the *sociometer theory*, which aims at explaining the function of self-esteem (Leary et al. 1995). A great deal of research has shown that individuals strive to feel good about themselves (i.e., achieve high self-esteem), yet high self-esteem does not seem to cause any benefits (Baumeister et al. 2003). Instead, according to the *sociometer theory*, self-esteem is a *result* of how much the individual is valued by others. Self-esteem therefore works as a kind of gauge, with high self-esteem reflecting high relational value and low self-esteem reflecting lack of social valuation. Experimental and longitudinal studies have provided support to the predictions of the *sociometer theory* (Anthony et al. 2007, Denissen et al. 2008). Therefore, rather than being an end in itself, the quest for high self-esteem ultimately represents a quest for social valuation. Analogously, the quest for a positive self-image can be interpreted as the quest for a positive social image.

Viewing the desired self-image as the desired reputation helps explain the cross-cultural variability in self-conceptions. There exist significant cross-cultural differences in how the self is constructed, the most documented of which is the difference between East Asian and North American selves (Heine 2001). North American culture is notoriously *individualistic*, with a strong emphasis placed on individual attributes and achievements as well as the respect of abstract, impartial rules. East Asian culture, on the other hand, is more *collectivist*, with a strong emphasis on personal relationships, inherited social roles and obligations. North Americans therefore

[4]Not everyone seeks to behave in accordance with what society values. In fact, individuals often actively seek to separate from the "majority" (Brewer 1991). Yet, as described by Smaldino (2019), the expression of an "unorthodox" social identity often has the function of finding similar others with whom to efficiently coordinate and cooperate. Therefore, these individual's ideal selves will be different from what (majority) society values and "designed" to their specific target audience.

face incentives to develop specific skills and abilities to be valuable as friends or partners; they also need to respect abstract, impartial rules and principles, in order to cultivate their reputation as trustworthy individuals (Henrich 2020). East Asians, on the other hand, face incentives to conform to in-group members, to favor the in-group over the out-group, to respect tradition and to defer to authorities (Heine 2001). It is therefore not surprising that self-enhancement, the motive to view oneself positively, is observed in North American but not in East Asian cultures (Heine & Hamamura 2007). North American selves are incentivized to enhance their skills and abilities, since this helps to appear as valuable and attract friends and partners. On the other hand, East Asians selves are more interdependent, focusing on their position in their social networks and the roles and obligations they need to fulfill (Heine 2001). The motive to improve one's self-image is therefore culture-specific. Moreover, given the strong emphasis on the cultivation of traits that identify individuals across contexts and relationships, North American selves seek self-consistency and abhor "cognitive dissonance". East Asian selves, on the other hand, are at ease with inconsistencies in behavior across contexts and relationships, if such inconsistency is expected from the specific roles and obligations that they need to fulfill in different situations (Heine 2001). The motive to remain self-consistent is therefore also culture-specific. It appears that North American and East Asian selves are tuned to their particular social environments; the *ideal selves* toward which individuals in North America and East Asia strive are a direct function of the specific cultural incentives that they face.

1.4 Learning and spillovers

A common argument, observed in a variety of scientific papers, is that the lack of an audience in laboratory experiments necessarily implies that individuals really care about their self-image when they, for instance, behave prosocially while anonymous. In their theoretical and experimental investigation of willful ignorance of information, Grossman & Van der Weele (2017) write that:

> Models that rely on concerns for reputation or social image cannot explain the decision to ignore information [...], as the experimental decisions were one-shot, anonymous and no participant observed whether or not the decision maker actually chose to be ignorant. (Grossman & Van der Weele 2017, p. 174)

Based on this argument, researchers have developed a variety of theoretical models (which will be discussed in Section 1.6) aimed at explaining behavior in lab-

oratory experiments, often referring to a self to which individuals must be signaling (Bodner & Prelec 2003, Bénabou & Tirole 2011, Grossman & Van der Weele 2017). Yet, another view, which will be introduced in this section, is that individuals learn (through reinforcement or social learning) what behaviors are optimal in their social environment, and these behaviors *spill over* when they find themselves in unusual and contrived environments, such as the laboratory.

The objective of this section is twofold. First, I want to describe the *reinforcement* learning and *social* learning frameworks and discuss why spillovers (or over-generalization) can be expected. Second, I want to provide evidence that such spillovers are real and widely documented.

1.4.1 How do we learn?

Humans are endowed with powerful learning mechanisms which allow them to adapt quickly to changing environments. *Reinforcement learning* (RL) problems capture situations in which learning occurs without the learner being told how to behave; the learner has to figure out what to do—through interaction with the environment—in order to maximize its future rewards (Sutton & Barto 2018). *Social learning* (SL), on the other hand, captures situations in which individuals learn from others, either by observing them, hearing about them, or through explicit teaching. Our SL capabilities have been said to underlie our extraordinary adaptation to very diverse environments (Boyd et al. 2011). Both types of learning work in concert so that individuals can adopt functional behaviors adapted to their environments.

In RL, two types of learning systems are distinguished. In *model-free* RL, the agent aims to find an optimal policy (i.e., an optimal action in each possible state) without learning a causal model of the environment. Rather, the agent updates the state-action values based on temporal difference errors: if the reward following a specific action in a specific state is higher than expected, then the value of that state-action is increased; conversely, if the reward is lower than expected, then the value is decreased.[5] This is computationally efficient, since the value of state-action pairs is updated locally based only on temporal difference errors. Nevertheless, *model-free* RL is inherently inflexible (it is thought to give rise to *habits*) and constantly requires trial-and-error learning in order to update the action-reward associations. In *model-based* RL, the agent aims to find an optimal policy by learning a causal model of the environment (i.e., by learning the transition and rewards functions of the Markov Decision Process). The knowledge of the causal model of the environment

[5] *Model-free* RL is therefore a form of Thorndike's Law of Effect, according to which actions that lead to rewards are more likely to be repeated (Thorndike 1927).

allows the agent to plan ahead (e.g., by dynamic programming) and to react flexibly to changes in the environment. *Model-based* RL therefore has the potential to be more accurate than *model-free* RL, but it is computationally expensive. Both RL systems compete (and cooperate) in control of behavior (Kool et al. 2018). In fact, there appears to be an intrinsic cost associated with exerting cognitive control, such that individuals tend to follow the "law of least mental effort" (Kool et al. 2010, Kool & Botvinick 2018). Yet, there also is evidence that individuals flexibly use both control systems as a function of the task at hand, and that the greater the incentives (expected reward) associated with the task, the greater the use of model-based control (Kool et al. 2017, Westbrook et al. 2013). Therefore, if the perceived stakes are low (which is typically the case in online or laboratory experiments), we might expect the (habitual) *model-free* system to control behavior.

As described above, individuals do not always have to rely on their own experience in order to learn from their environment. Humans are also endowed with powerful SL mechanisms which allow them to pick useful information from others. These mechanisms appear to be sensitive to a variety of factors, such as the number, confidence and competence of others, the task difficulty, the reliability of one's own information, or the cost of individual learning, suggesting that SL mechanisms are governed by adaptive rules (Morgan et al. 2012, Toelch et al. 2014). Research with young infants has shown that these adaptive learning mechanisms develop early on (Birch et al. 2008, Corriveau & Harris 2009), with infants already discriminating between competent and incompetent informants. Yet, while our SL mechanisms appear to be finely tuned and context-sensitive, researchers have also documented cases of *overimitation*, which refers to the copying of arbitrary and unnecessary actions (McGuigan et al. 2011). While seemingly inefficient, researchers have argued that overimitation might be a feature, rather than a bug: given the intensive reliance on cultural information for survival and given the often causally opaque nature of such information, overimitation might be a fast and adaptive way of acquiring useful skills or knowledge (Boyd et al. 2011, McGuigan et al. 2011). Similarly, given our history of living in groups governed by social norms (shared and enforced behavioral standards), researchers have argued that we are endowed with a *norm-psychology*, a set of psychological mechanisms suited for acquiring and internalizing social norms (Chudek & Henrich 2011). The internalization of social norms is thought to prevent defection and—as a consequence—help preserve one's reputation. Therefore, while our SL mechanisms appear sophisticated, fine-tuned, and task- and context-specific, SL theorists also argue that the complexity of cultural information, associated with the importance of recognizing and adopting social norms, might have led to the evo-

lution of mechanisms whose function is to acquire rapidly—and internalize—certain kinds of information.

Consequently, both *model-free* (habitual) control of behavior and internalization predict spillovers in the laboratory. The next section will be dedicated at rapidly reviewing the large literature documenting such spillovers.[6]

1.4.2 Documented spillovers

Cross-cultural variation in cooperative behavior

The first evidence for spillovers from everyday life to the laboratory comes from cross-cultural studies of behavior in economic experiments, such as the dictator, ultimatum and public good games. In their seminal paper, Henrich et al. (2001) have shown that there exist large cross-cultural variations in how individuals behave when playing such games in the laboratory. Importantly, this cross-cultural variability can be traced directly to differences in economic organization and degree of market integration across societies: "the higher the degree of market integration and the higher the payoffs to cooperation, the greater the level of cooperation in experimental games" (Henrich et al. 2001, p.74). This team of researchers has concluded their cross-cultural study by noting that behavior in the laboratory is consistent with patterns of behavior in everyday life. In a telling excerpt, they write:

> The Machiguenga show the lowest cooperation rates in public-good games, reflecting ethnographic descriptions of Machiguenga life, which report little cooperation, exchange, or sharing beyond the family unit. By contrast, Orma experimental subjects quickly dubbed the public-goods experiment a harambee game, referring to the widespread institution of village-level voluntary contributions for public-goods projects such as schools or roads. Not surprisingly, they contributed generously (58 percent of the stake), somewhat higher than most U.S. subjects contribute in similar experiments. (Henrich et al. 2001, p.76)

Since then, other researchers have undertaken the task of comparing behavior in economic experiments across societies and the results support the idea that behavior in the laboratory reflects learned behavior outside the laboratory (Gächter et al. 2010, Herrmann et al. 2008).

[6] See the Appendix for further evidence of such spillovers in the context of Classroom Experiments with students.

Social heuristics

According to the *Social Heuristics Hypothesis* (SHH), "people internalize strategies that are typically advantageous and successful in their daily social interactions" (Rand, Peysakhovich, Kraft-Todd, Newman, Wurzbacher, Nowak & Greene 2014, p.2), and they bring these strategies (as heuristics) in the laboratory. If true, then we expect that cooperation would be intuitive only for those individuals who have cooperative relationships outside the laboratory. This is, in fact, what researchers have found: it is only for individuals who report having cooperative relationships outside the laboratory that cooperation is intuitive (Rand et al. 2012, Rand, Peysakhovich, Kraft-Todd, Newman, Wurzbacher, Nowak & Greene 2014). Interestingly, when intuitive cooperators have time to deliberate (i.e., have time to think about the novel settings) cooperation is reduced, suggesting a transition from *model-free* to *model-based* control of behavior (Rand et al. 2012, Rand & Kraft-Todd 2014). Moreover, by experimentally exposing individuals to environments in which cooperation is either favored or not, Peysakhovich & Rand (2016) are able to show that individuals that are exposed to environments supportive of cooperation are more likely to become intuitive cooperators in subsequent one-shot interactions. Similarly, Stagnaro et al. (2017) experimentally show that the quality of institutions causally influences subsequent behavior in one-shot interactions, with institutions promoting cooperation leading to increased prosociality in one-shot games. This reliance on intuitive cooperation in one-shot interactions, when individuals usually have repeated cooperative relationships and deliberation is costly, has been shown to emerge at equilibrium in evolutionary game-theoretic models of the evolution of cooperation (Bear & Rand 2016). This work therefore provides additional evidence that subjects in experimentally induced one-shot interactions use heuristics that they have internalized from everyday interactions outside the laboratory.

Learning and reasoning in games

If spillovers from everyday interactions underlie behavior in the laboratory, then we expect individuals to update their behavior once they get accustomed to the new, contrived experimental settings. In fact, there is considerable evidence that subjects used to playing games in laboratory settings converge over time to the optimal strategy. For instance, Conte et al. (2019) show that contributions in the public goods game decrease with experience of the game, suggesting that individuals have learned the payoff-maximizing strategy. Similarly, McAuliffe et al. (2018) show that individuals in economic games become less cooperative over time when cooperation can not promote self-interest; yet, subjects continue cooperating when cooperation

is payoff-maximizing, reflecting their learning of the optimal strategy. In a meta-analysis of the public goods game literature, Burton-Chellew & West (2021) conclude that *payoff-based learning* underlies the decline in cooperation over time, since decreased cooperation is faster when individuals have more influence over their payoffs, thereby facilitating learning. Finally, greater reliance on the *model-based* system also reduces cooperation in economic games. For instance, Burton-Chellew et al. (2016) show that subjects who have a better grasp of the rules of the game cooperate less, while Barreda-Tarrazona et al. (2017) show that reasoning ability is associated with a decreased probability of playing cooperatively.

1.5 Game-theoretic intuitions

Before reviewing the self-signaling framework in light of the arguments exposed in this chapter, I will show in this section that more than just behavior, learning processes tracking rewards and punishments can also shape our beliefs, preferences and intuitions. To illustrate this, I will discuss three examples, the *omission-commission distinction*, *ineffective altruism* and our *sense of rights*, and describe how fine-tuned to social incentives our proximate mechanisms can be.[7]

1.5.1 The omission-commission distinction

People usually feel (or intuit) that engaging in moral violations by omission is more acceptable than by commission, even though the end result is typically the same. Similarly, observers tend to judge acts of omission less harshly than acts of commission. For instance, Spranca et al. (1991) show that subjects judge that willingly poisoning someone (an act of commission) is typically worse than withholding information about the presence of poison (an act of omission), even though both actions lead to the same outcome. While we might think of this distinction as a simple bias, DeScioli et al. (2011) have proposed that the preference for omissions is strategic: people choose omissions in order to avoid being punished (morally condemned) by others. They provide evidence for this claim by showing that the frequency of omission increases when punishment by third-parties is possible. Yet, why might omissions, compared to commissions, remain unpunished? Moral punishment is typically coordinated, so that third-parties would like to punish a moral violation only if they expect others to punish too (DeScioli & Kurzban 2013). In a game-theoretic

[7] An extensive treatment of how game-theoretical tools can help us better understand the beliefs and preferences that individuals adopt can be found in Hoffman et al. (2016) and Hoffman & Yoeli (2022).

analysis of coordinated enforcement, Hoffman et al. (2018) show that third-party coordination is easiest when transgressions are sufficiently observable (the signal is public) and shared (the signal is correlated). Commissions (actions) are typically observable and can therefore trigger coordinated punishment. On the other hand, omissions (intentions) are not easily observable since they typically leave room for plausible deniability.[8] The argument is therefore that people *feel* that omissions are more acceptable than commissions since they typically remain unpunished, while third-parties *judge* omissions less harshly since they do not generate enough common knowledge of transgression among observers.

1.5.2 Ineffective altruism

People are motivated to give, but not to give effectively. This failure to take efficacy into account has often been interpreted as stemming from cognitive or emotional limitations. Yet, the ineffectiveness of giving is predicted by taking a closer look at evolutionary game-theory models of cooperation. As described by Burum et al. (2020):

> In all such models, the following criteria are crucial for maintaining good behaviour: what counts as good must be (1) well defined and (2) easy to observe, and (3) different people's assessments must be correlated with each other or easily communicated—that is, they must create common knowledge. (Burum et al. 2020, p.1245)

The issue with the effectiveness of giving is that effectiveness is not easily defined nor easily measured. This implies that third-parties might not be able to reward cooperators based on the efficacy of their act. However, the *act* of cooperating (or giving) is usually well defined and easily communicated to others. In support of this account of ineffective giving, Burum et al. (2020) show in a series of experiments that individuals are capable of taking efficacy into account when making savings (but not charitable-giving) decisions, or when giving to a kin instead of a stranger, suggesting that cognitive limitations are not the root cause of ineffective altruism. Additionally, they show that third-parties typically reward others on the basis of their giving but not on the effectiveness of the gift. Since social rewards depend on whether or not we give but not on the effectiveness of our gifts, we have no incentive to take efficacy into account. Our proximate mechanisms are well-attuned

[8] In the above example, the person withholding information about the presence of poison can always argue that she did not know about the presence of poison, while the *act* of poisoning leaves little room for doubt.

to such incentives. For instance, when observability is increased, contributions tend to increase dramatically (Alpizar et al. 2008, Funk 2010, Yoeli et al. 2013). On the other hand, individuals go to great lengths to avoid situations in which they might be asked to contribute and contribute only if they have no excuse not to (Andreoni et al. 2017). What Andreoni (1990) describes as the *warm-glow* of giving therefore appears to be context-specific, and most importantly, responsive to social incentives.

1.5.3 Sense of rights

Individuals have strong intuitions about property. For instance, *first possession* strongly influences judgments of ownership in adults and infants (Friedman 2008, Friedman & Neary 2008), while *investment of creative labor* in a property is typically judged as transferring ownership to the person making the investment (Kanngiesser et al. 2010). In a virtual environment in which subjects could contest for berry patches (later convertible to cash), DeScioli & Wilson (2011) show that *who arrived first* at the patch determined the outcome of the contest more often than fighting ability. Arriving first (or first possessing the object) does not alter the payoffs of the game nor the probability of success in fighting, so why do we condition our sense of ownership and property on such seemingly arbitrary features?

In an extension of the classic Hawk-Dove game, Maynard Smith (1982) has shown that the Bourgeois strategy, which plays Hawk when owner but Dove when intruder, is the only evolutionary stable strategy of the game. That is, conditioning play in the contest for a resource on an *uncorrelated asymmetry*—here, who arrived first—is a central prediction of the extended Hawk-Dove game. Uncorrelated asymmetries set expectations about the behavior of the other player and can thus allow players to avoid the costs of fighting, particularly when such costs are large (relative to the value of the resource) and when fighting ability is roughly equal among players (Gintis 2007).

Animal studies have provided support to the predictions of the extended H-D game, since the "prior-residence effect" has been observed in a variety of species (Kokko et al. 2006). Another testable prediction of the extended H-D game is that disagreement about ownership is likely to occur when the players do not agree about which uncorrelated asymmetry to condition on. In a series of vignette experiments based on classical property law cases, DeScioli & Karpoff (2015) show that disagreement about ownership among subjects is highest when two uncorrelated asymmetries conflict, such as whether an object belongs to the finder ("finders, keepers") or to the person owning the land on which the object was found.

Therefore, it appears that our (and other animals) intuitions about ownership

and property follow the logic of the extended H-D game, in the sense that our beliefs about who owns what are shaped by strategic considerations related to the acquisition of valuable resources and the avoidance of costly fighting.

1.6 Self-signaling revisited

According to the self-signaling framework, individuals are uncertain about their true type (or their dispositions), such as whether they are intrinsically good, moral, honest, healthy, or clever, but they can learn about their type from their behavior (Bodner & Prelec 2003, Bénabou & Tirole 2004, 2011). More specifically, individuals in self-signaling models are typically divided in two: there is the *decision-making* self and the *observer* self. The decision-making self is assumed to know their true type and acts accordingly. The observer self lacks introspective knowledge of their type and can only infer their type from the behavior of the decision-making self. The inference individuals (observers) make about their type from their own behavior provides them with *diagnostic utility* (Bodner & Prelec 2003), which can be interpreted as benefits from a positive self-image. Since they derive (diagnostic) utility from thinking highly about themselves, the main idea is that individuals (decision-makers) will behave in such a way as to signal to themselves (observers) that they are high types, in the sense that the decision-making self will manufacture "diagnostic" signals about their type which will then be interpreted impartially by the observer self (Bénabou 2015).

The typical decision-making environment, taken from Bodner & Prelec (2003), is as follows. Let X represent the set of possible actions, with $x \in X$ denoting a specific action. Let θ represent the individual's type and $f(\theta)$ the individual's prior belief (distribution) about their type. Finally, $V(\theta)$ is a meta-utility function which represents the individual's preferences over their type. The total utility U of an individual choosing $x \in X$ is written as follows:

$$U(x, X, \theta) = u(x, \theta) + \sum_{\theta} f(\theta|x, X)V(\theta).$$

The first term of the total utility, $u(x, \theta)$, represents the material utility of choosing x when the type is θ. The material utility depends on θ, not its expectation, given that the decision-making self is supposed to know their type. The second term, $\sum_{\theta} f(\theta|x, X)V(\theta)$, represents the diagnostic utility, with $f(\theta|x, X)$ the *interpretation function* updating the self-image in light of the chosen action. Since $V(\theta)$ is typically increasing with θ, individuals are motivated to interpret their behavior in such a way

as high values of θ have a high probability.

The inference individuals can make about their type from their own behavior is not without constraints, however. It is typically assumed that (observer) interpretations about the individual's type need to be rational, updated using Bayes' Rule (Bodner & Prelec 2003, Bénabou & Tirole 2011, Grossman & Van der Weele 2017). Therefore, individuals can not just fool themselves into believing that they are the highest possible type. Their beliefs must be consistent with the way they behave. According to the self-signaling framework, this consistency constraint underlies puzzling behavior in the laboratory, such as ignorance of information about the social impact of one's action, when such ignorance can serve to obfuscate (in the eyes of the observer) the choice the individual (decision-maker) would have made had they received the information (Grossman & Van der Weele 2017).

Self-signaling interpretations therefore require that we accept that (i) individuals at times know their true preferences but at other times can not sufficiently introspect, that (ii) they can manufacture "diagnostic" signals which they then interpret as impartial, and that (iii) a positive self-image is a fundamental motive which warrants such mind games. Given the arguments developed in this chapter, I believe that the claim that laboratory experiments tap into a psychology well-adapted to the social incentives of everyday life is more parsimonious. In fact, Grossman & Van der Weele (2017, p.177) note that self- and social-signaling models are "technically equivalent", yet they reject the social-signaling interpretation since interactions in the laboratory are one-shot and anonymous. Given the evidence reviewed in Section 1.4 and Section 1.5, this argument appears to lose its bite.

As a matter of example, *lying aversion*—the fact that individuals do not always lie in the laboratory when it is in their self-interest to do so—has often been interpreted through the lens of the self-signaling model. Researchers have argued that, since subjects are anonymous and interact one-shot, the prevalence of lying aversion must stem from psychological costs associated with lying or from a disutility from thinking of oneself as a bad person. Yet, lying aversion, just as cooperative behavior reviewed in Section 1.4.2, likely stems from everyday life useful heuristics spilling over in the laboratory. Lying can be costly if spotted, since the liar might not be trusted anymore (Vullioud et al. 2017). In environments in which lies are punished and honesty is the norm, we therefore expect individuals to display an "intrinsic" aversion to lying. In a cross-cultural study from 23 countries, Gächter & Schulz (2016) show that "intrinsic" honesty, measured in anonymous die-rolling experiments, closely tracks an index of society-wide prevalence of rule violations. As expected, the greater the prevalence of rule violations in a society (which cor-

relates with a lack of enforcement of rules and norms), the smallest the "intrinsic" honesty that individuals display in the laboratory. That is, when dishonesty is the norm outside the laboratory, individuals do not tend to display an "intrinsic" aversion to lie in the laboratory. Additionally, if lying aversion is a learned strategy to avoid punishment and reputational damage, we expect individuals to modulate their aversion to lie as a function of the probability of being caught. Several recent experiments have in fact shown that subjects lie less, the greater the observability of their lie (Fries et al. 2021, Gneezy et al. 2018), or the smaller the plausible deniability (Shalvi et al. 2011). A final piece of evidence against a self-signaling motive in lying aversion comes from Bašić & Quercia (2022), who show that manipulating self-awareness in the laboratory has no effect on overreporting in die-rolling experiments. On the other hand, manipulating the social image of subjects (with respect to the experimenter) significantly reduces lying, in line with the idea that individuals are motivated to maintain their reputation as honest and trustworthy.

1.7 Implications

Self-signaling models suggest that individuals are primarily motivated to appear good to themselves and that they behave in such a way as to convince themselves that they are, in fact, good, fair or honest. This is a fundamentally individualistic perspective that is at odds with a social-signaling perspective. In fact, the self-signaling perspective does not provide any clear guidelines into how policymakers might promote individually-costly prosocial behavior, such as contribution to public goods or honesty. One might think that policymakers could promote prosocial behavior by framing the decision-problem so as to appeal to an individual's self-image or self-signaling concerns. But since self-signaling is fundamentally individualistic, any individual might—*a priori*—have their own particular idea of what is good, fair or honest. The particular framing chosen by the policymaker is therefore unlikely to be aligned with the preferences of the population and, therefore, unlikely to appeal to the population's self-image or self-signaling concerns. As a matter of fact, the observation that individual values (what is good, fair, honest, etc.) are highly correlated among individuals belonging to the same group/culture/society strongly suggests that such values are socially influenced/enforced (Henrich 2020), which provides additional support to the social-signaling perspective defended in this chapter.

According to the arguments developed in this chapter, individuals are not (ultimately) motivated to self-signal or improve their self-image. Instead, what matters

is their social image (or reputation), since how others perceive them has important material and social consequences. Individuals therefore develop a variety of strategies to navigate their social environments and maintain their reputation. These strategies then tend to spill over when they find themselves in anonymous or private settings. It is actually a central prediction of evolutionary theory that behavior (and therefore proximate psychological mechanisms more generally) will be adapted only to the environment(s) in which it was selected for (West et al. 2011). The cross-cultural studies reviewed in this chapter provide strong support for such a perspective. Everywhere around the world, individuals adapt their behavior (and proximate psychological mechanisms) to the particular social environment in which they find themselves. Depending on variables such as market integration (Henrich et al. 2010), prevalence of rule violation (Gächter & Schulz 2016), state centralization (Lowes et al. 2017) or kinship intensity (Schulz et al. 2019), which all modify the incentives that individuals face in their everyday lives, significant behavioral differences in the laboratory are being recorded. More than just behavior, our preferences (Billing & Sherman 1998), beliefs (Henrich & Henrich 2010) or thinking styles (Nisbett et al. 2001) also become adapted to our social environments (see also Section 1.5).

These observations point to the crucial role of the social environment in shaping our behavior and psychology. This suggests that to promote individually-costly prosocial behavior, systemic-level interventions are needed (Chater & Loewenstein 2022). Given the importance of formal and informal institutions (social norms) in regulating social interactions and defining the payoffs for different behaviors (Powers et al. 2016), policymakers should primarily aim at developing and stabilizing institutions and organizations incentivizing prosocial behavior. Experimental work has already shown that exposure to structures incentivizing cooperation improves cooperative outcomes, even in subsequent one-shot interactions (Peysakhovich & Rand 2016, Stagnaro et al. 2017). The argument is that repeated exposure to such institutions (or incentives to cooperate) will make cooperation "intuitive" and "the right thing to do" through the adaptation of proximate psychological mechanisms (beliefs, preferences, intuitions) to the social environment.

1.8 Conclusion

This chapter has been concerned with reevaluating the idea that (i) individuals (ultimately) care about improving their *self*-image and (ii) individuals *self*-signal. This has been done in several steps. First, evolution only cares about outcomes,

which is why improving one's self-image or convincing oneself of something for its own sake, can not be goals our minds are designed to achieve. Our mind games must necessarily have an effect outside our bodies. Moreover, rather than being incentives or rewards, feelings typically constitute reflections about (i) our current condition and/or (ii) the need to maintain or improve our current state.

Second, the self is not a "thing" which *has* all the experiences of life. There is no entity that is us, different from what we do, say or think. Rather, the self is the collection of all experiences of life (our thoughts, memories, desires and sensations) all integrated and unified in the mind. Being self-aware therefore implies turning one's attention and thoughts to one's actions, desires, memories and experiences. Since wanting to improve one's self-image can not be an end in itself, the desired self-image is best seen as the desired reputation: individuals will feel good about how they see themselves (i.e., their self-image) when their behavior is in accordance with how they want to be known by others (i.e., their desired reputation).

Third, both the reinforcement learning and the social learning frameworks predict spillovers from everyday life to the laboratory if (i) behavior is controlled by the model-free (habitual) system and/or (ii) codes of conduct have become internalized. The existing empirical evidence is in line with the idea that behavior in the laboratory reflects patterns of behavior in everyday life, such that individuals bring in the laboratory strategies (heuristics) that have proved useful outside the laboratory. Further evidence for the importance of spillovers is that experience with experimental settings lead individuals to converge to the optimal strategy over time.

Fourth, more than just behavior, learning processes tracking rewards and punishments also shape our beliefs, preferences and intuitions. Given these arguments, the claim is the following: what appears to be self-signaling in the laboratory reflects the workings of a psychology well-tuned to the social incentives of everyday life, which spills over when individuals find themselves in unfamiliar environments, such as the laboratory. This claim is more parsimonious than the self-signaling interpretation, which requires that we accept that (i) individuals at times know their true preferences but at other times do not, (ii) they behave in such a way as to convince themselves and then interpret their behavior as impartial, and (iii) a positive self-image is a fundamental motive.

This changing perspective, from the *self* to the social, has important practical and policy relevance. The cross-cultural studies discussed in this chapter show that the way individuals behave anonymously tends to reflect the way they behave in their everyday lives. Similarly, the intuitions individuals develop when interacting with others tend to spillover in the laboratory, even if they are by themselves. This

is apparent in findings that lying tends to be reduced, the greater the observability (i.e., the greater the probability of being caught) of the lie. If the objective is to promote individually-costly prosocial behavior, then appeals to self-image are likely to be insufficient. Rather, the development of institutions and organizations that incentivize cooperative behavior—either by rewarding cooperation or punishing selfish behavior—is needed, since repeated interactions with such structures have been shown to promote cooperative behavior even in one-shot (anonymous) interactions.

1.9 Appendix: Classroom experiments

This section will further illustrate the argument that individuals bring into the laboratory heuristics (or intuitions) that they have developed outside the laboratory. To that effect, I will analyze data collected from Classroom Experiments by Gisèle Umbhauer in September and October 2022. The games were played by Undergraduate (third-year) students at the Faculty of Economics and Management of Strasbourg (France) enrolled in Gisèle Umbhauer's "Games and Strategies" course. The games were played in the classroom during lecture hours and students had to make their decision individually on an online platform. The instructions were publicly explained before each game. Students were free to decide not the play the games and they were invited to explain their choices in writing. Students were not monetarily incentivized but were told that participating in the experiments could improve their final grade.

1.9.1 Gift-Exchange game with "employer-employee" context

The first game played was a "gift-exchange" (or "employer-employee") game (see Figure 1.1). The term "gift-exchange" was not used to describe the game so as to avoid students being nudged towards one specific way of playing. The game was played by 240 students. P_1 is considered to be the employer who has to decide which *wage* she will offer to the employee P_2. She has to decide between three wages W_S, W_M and W_L, with $W_S < W_M < W_L$. Once the employer has decided which wage to offer, the employee (P_2) has to decide which level of *effort* to exert between E_S, E_M or E_L, with $E_S < E_M < E_L$. The employer's payoffs are displayed first in the list. The payoffs are such that:

— The greater W is, the greater the employee (P_2)'s payoff is: for a given amount of effort, his payoff increases with the wage offered.

— The greater E is, the lower the employee (P_2)'s payoff is: for a given wage, his utility decreases with the amount of effort exerted.

— The greater W is, the lower the employer (P_1)'s payoff is: for a given amount of employee effort, her payoff decreases with the amount of wage offered.

— The greater E is, the greater the employer (P_1)'s payoff is: for a given amount of wage offered, her payoff increases with the amount of effort exerted.

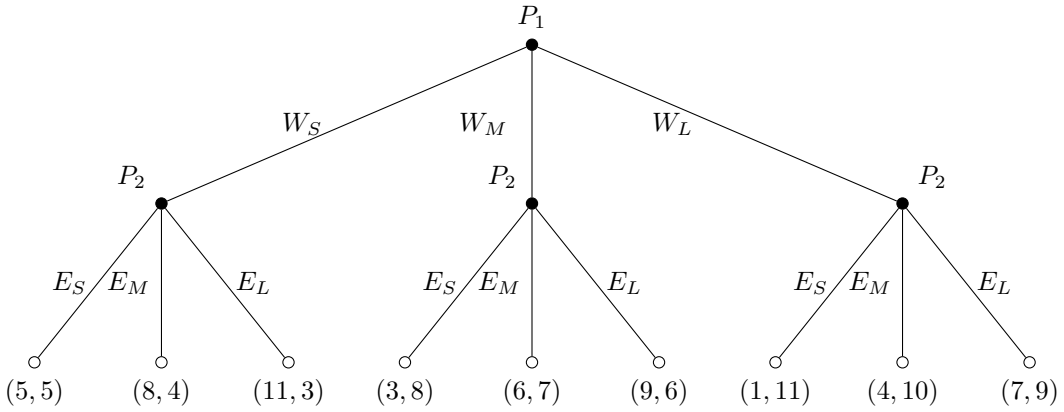


Figure 1.1 – Gift-Exchange game with "employer-employee" context.

Students were in the role of the employee (P_2) and their only decision was, therefore, to choose their effort level (E) for every wage offered. The game was played only once. The employee's best response is to play E_S (that is, provide the least effort) for every wage offered by the employer. The only Subgame-Perfect Nash Equilibrium of this game is therefore $\{W_S, E_S\}$ at which the employer (P_1) offers the lowest wage and the employee (P_2) exerts minimum effort.

The results of this game are shown in Figure 1.2, together with the results of Pearson's Chi-squared test of independence between the *wage* offered and the *effort* exerted. One can see that when students face a low wage offer W_S, almost all of them (around 92%) play the best response E_S, exerting minimal effort. Yet, the greater the offered wage, the smaller the fraction of students playing the best response E_S. When the employer offers the medium wage W_M, only around 69% of the students decide to play E_S, with around 27% of students deciding to exert medium effort E_M. When the employer offers the highest possible wage W_L, only around 63% of students play E_S, with around 20% of them exerting maximal effort, even though E_L generates the smallest payoff for employees. The *p-value* of the Chi-squared test of independence being strictly lower than 0.01 (p=1.78e-20), we can reject the null hypothesis that both the *wage* offered and the *effort* exerted

are independent variables. Interestingly, we can observe from the students' written transcripts that those that exert a medium or high effort often do not look at the employer's payoff. Rather, they justify their decision by using maxims such as "effort must be proportional to the offered wage" or "one must behave honestly towards one's employer". This suggests that (i) (some) students have in some way "internalized" a reciprocity norm which specifies that favours or "gifts" (here, a medium or high wage offer) need to be returned (here, by exerting medium or high effort) and/or (ii) (some) students judge that exerting low effort after having been proposed a medium or high wage would be considered "cheating" and therefore refrain from doing so.

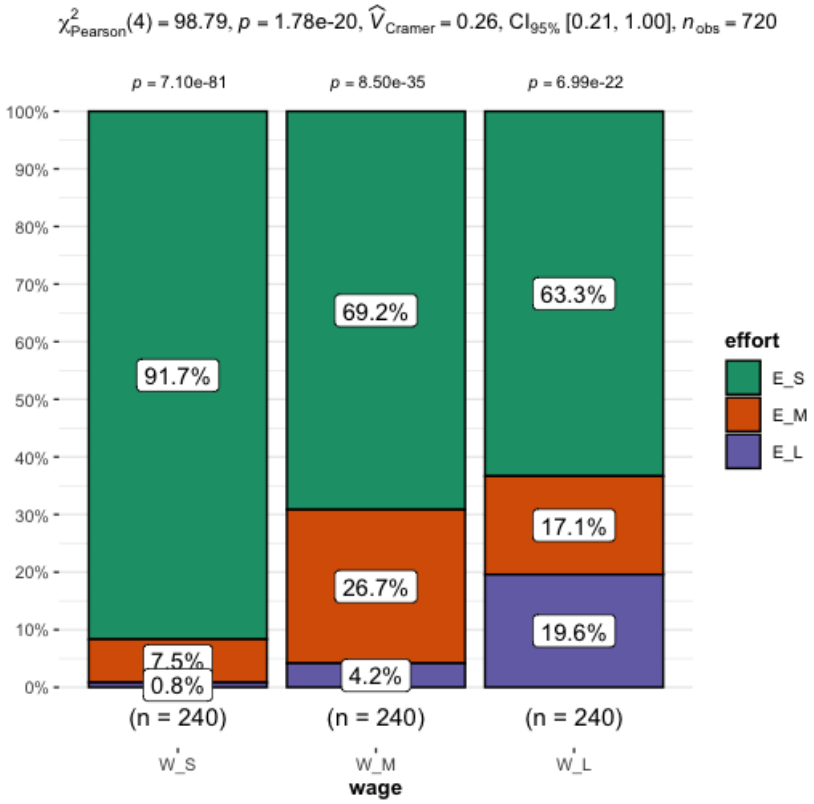


Figure 1.2 – Game "employer-employee" results.

1.9.2 Gift-Exchange game without context

Students also played a variant of the gift-exchange game in which the "employer-employee" context has been stripped away. More specifically, they played the abstract games G_1 (Figure 1.3) and G_2 (Figure 1.4). Game G_1 and Game G_2 were played by 190 students and were played one week after the "employer-employee" game so as to prevent (as far as possible) students from linking the G_1 and G_2

games with the "employer-employee" game.

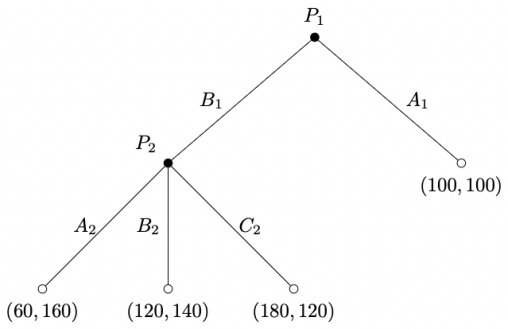


Figure 1.3 – Game G_1.

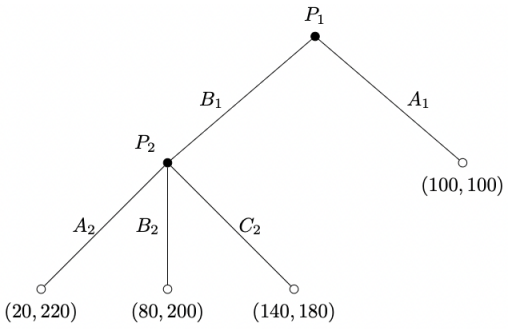


Figure 1.4 – Game G_2.

G_1 and G_2 are variants of the gift-exchange game in that A_1 can be considered as Player 1 (P_1)'s "subgame-perfect" strategy (since one week before around 92% of students exerted minimal effort after being offered a low wage W_S), while B_1 in game G_1 can be considered as the medium wage (W_M) offer and B_1 in game G_2 the high wage (W_L) offer. The payoffs have been multiplied by 20 compared to the "employer-employee" game, but the incentive-structure, of course, remains the same. In both games, students were in the role of Player 2 (P_2) and were asked to decide what to play (A_2, B_2 or C_2, which can be considered as effort levels) at their decision node. As in the gift-exchange game with the "employer-employee" context, P_2's best response is always to play A_2 (which can be considered as exerting the lowest effort).

The results are shown in Figure 1.5 (note that we have kept the previous action names in order to facilitate comparison). We can observe that when the context is stripped away, almost all students (90% in G_1 and around 87% in G_2) play the best response A_2. The *p-value* of the Chi-squared test of independence between G_1 and G_2 is exactly equal to 0.05, suggesting that the game played did not significantly influence the actions students decided to play. Conversely, the Chi-squared test of independence between the W_M and W_L treatments in the "employer-employee" context generates a *p-value* equal to 3.60e-07 (see Figure 1.7), confirming that the context has a significant influence on how the students decide to play the game. Therefore, we can conclude that when the "employer-employee" context is removed, the "internalized" reciprocity norm is not activated and almost all students choose to maximize their own payoff.

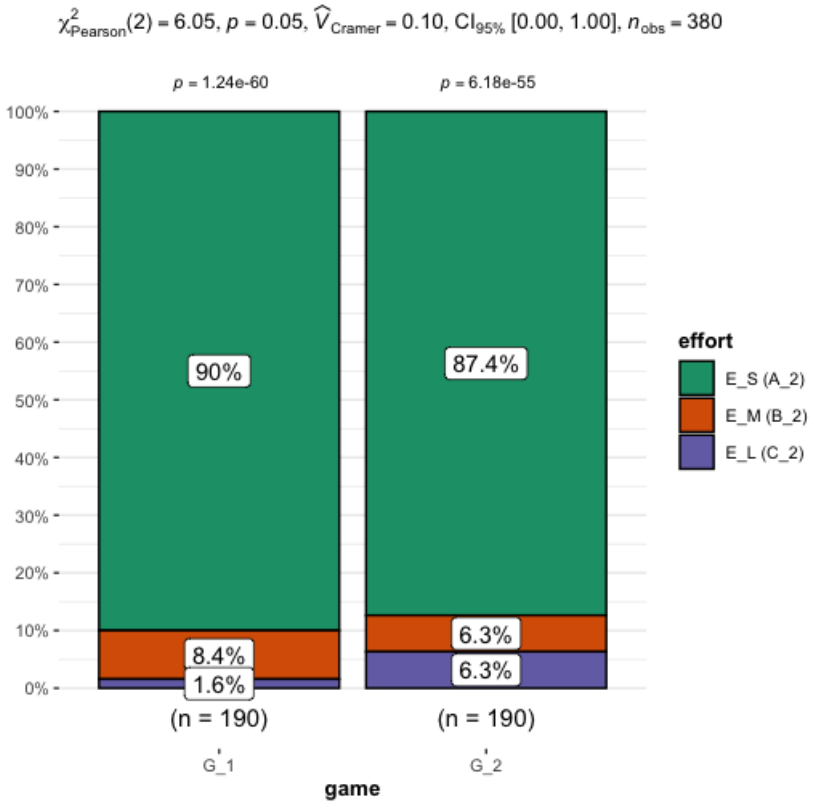


Figure 1.5 – Games G_1 and G_2 results.

1.9.3 Gift-Exchange game without context but with communication

In the last variant, 190 students played the games G_1 (Figure 1.3) and G_2 (Figure 1.4) but this time with a communication round before the game started. More specifically, before choosing their action at their decision node, students received the following message from Player 1 (P_1):

> I could have played A_1, giving you a payoff of 100 and ensuring myself a payoff of 100. By playing B_1, I allow both of us to have a payoff strictly greater than 100, provided that you play a certain way. It's your turn to play.

Therefore, while the "employer-employee" context is still stripped away, the abstract G_1 and G_2 games gain meaning once communication is possible. By emphasizing that playing B_1 can allow both players to earn a greater payoff and by trusting P_2 to "do the right thing", P_1's message certainly taps into intuitions or heuristics that students have learned or acquired in everyday life. For one, P_1 explicitly emphasizes her expectation of reciprocity from the part of P_2. Second, by

stressing that *both* players can have a greater payoff, P_1's message generates some kind of *shared* goal. Finally, by trusting P_2 not to play selfishly, P_1's message without a doubt taps into internalized norms of justice and/or fairness.

The results are shown in Figure 1.6. As expected, the communication round had a significant influence on the students' decisions. In G_1, only around 24% of students played the best response A_2, while around 71% of them "reciprocated" the favour by choosing the action B_2 that increases both players' payoffs (compared to A_1) *and* minimizes the payoff difference between both players. In G_2, only around 22% of students decided to play A_2, while around 71% of them played C_2, which is the only action that increases both players' payoff. The *p-value* of the Chi-squared test of independence being strictly lower than 0.01 (p=8.58e-46), we can reject the null hypothesis of independence between the G_1 and G_2 games with communication.

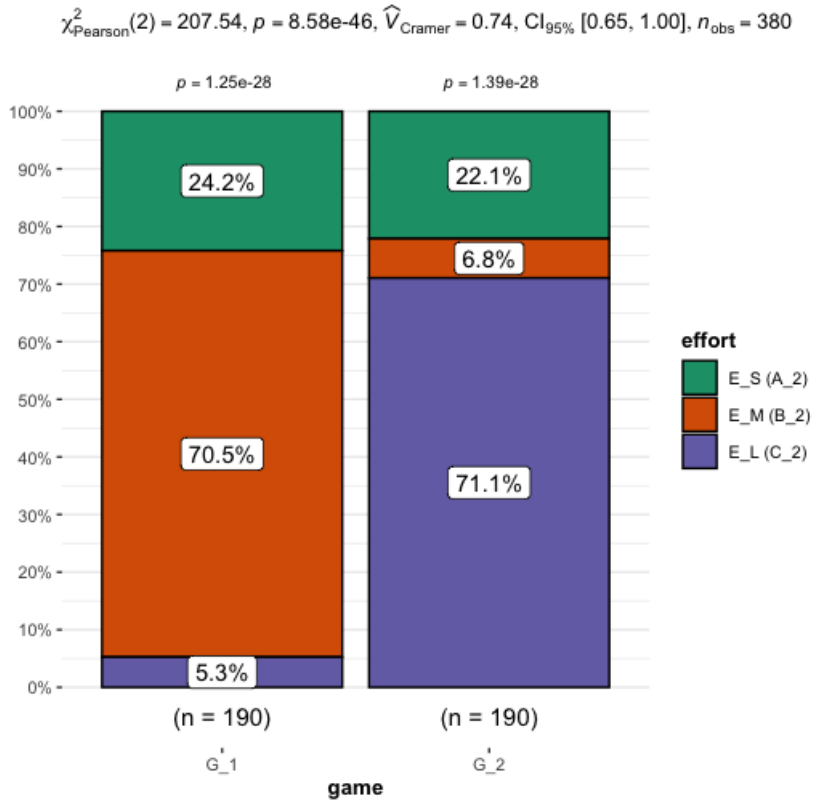


Figure 1.6 – Games G_1 and G_2 results (with communication).

1.9.4 Discussion

Figure 1.7 presents the statistic and p-value of Pearson's Chi-Square test of independence *across* the relevant treatments (G_1 vs W_M, G_2 vs W_L, G_1(com) vs W_M, G_2(com) vs W_L, G_1 vs G_1(com) and G_2 vs G_2(com)). We can observe that the

Treatment	G_1	G_1(com)	G_2	G_2(com)	W_M	W_L
G_1	-	168.6	6.05	-	27.2	-
(p-value)	-	(2.45e-37)	(0.05)	-	(1.24e-06)	-
G_1(com)	168.6	-	-	-	88.04	-
(p-value)	(2.45e-37)	-	-	-	(7.59e-20)	-
G_2	6.05	-	-	176.88	-	31.86
(p-value)	(0.05)	-	-	(3.89e-39)	-	(1.20e-07)
G_2(com)	-	-	176.88	-	-	115.18
(p-value)	-	-	(3.89e-39)	-	-	(9.73e-26)
W_M	27.2	88.04	-	-	-	29.67
(p-value)	(1.24e-06)	(7.59e-20)	-	-	-	(3.60e-07)
W_L	-	-	31.86	115.18	29.67	-
(p-value)	-	-	(1.20e-07)	(9.73e-26)	(3.60e-07)	-

Figure 1.7 – Statistic and p-value of Pearson's Chi-Square test of independence across treatments.

students' choices were significantly different across all treatments (except G_1 vs G_2 as discussed above), meaning that the difference in context significantly influenced their way of playing.

The results suggest that the game's framing has important implications about how students decide to behave. Across all treatments, the incentive-structure remains constant and yet we can observe significant differences in how students choose between their three actions. The argument is that this "framing effect" arises from useful heuristics/intuitions (reciprocity, justice, fairness, etc.) developed in everyday life, influencing how students decide in the classroom.

Chapter 2

An Evolutionary Perspective on Social Preferences

Summary

In this chapter, I argue that the theory of social evolution has the necessary scope and power to provide a useful theoretical framework for human social preferences. Social evolution theory applies to any social interaction and makes sharp predictions about the kinds of social traits that can be observed in the population. It illuminates the function of our social emotions, which are context-dependent, endogenous mechanisms that modulate our social relationships with others. I discuss two mechanisms that underlie the wide variation in the expression of social preferences: social norms (or institutions) and local ecology. The wealth of findings on human social preferences supports their context-dependent nature and prompts us to investigate the incentives underlying their expression.

Classification

JEL Classification: C70, D91

Keywords: Behavioral Ecology, Institutions, Social Emotions, Social Evolution, Social Preferences

2.1 Introduction

The recent decades have observed a surge in findings describing how human behavior consistently differs from the predictions of traditional game theory and standard economic theory. These findings have led to the emergence of a new research program—known as *behavioral game theory*—which "aims to replace descriptively inaccurate modelling principles with more psychologically reasonable ones" (Camerer 1997, p.185). In this chapter, I will concentrate on the finding that, contrary to what has been suggested by the traditional view, humans are not solely self-interested but are endowed with *other-regarding* (or *social*) preferences.

Starting in the eighties, simple games have been used to test the predictions of traditional game theory. A common finding from these experiments is that subjects in the laboratory fail to play according to traditional equilibrium predictions. Instead, individuals appear to be concerned with *fairness* (Güth et al. 1982) or *inequity-aversion* (Fehr & Schmidt 1999), *envy* (Kirchsteiger 1994), *trust* and *reciprocity* (Berg et al. 1995), or *altruism* and *spitefulness* (Andreoni & Miller 1993, Levine 1998); in other words, humans appear to have *social preferences*. In an effort to upgrade game-theoretic models in light of this new wave of evidence, researchers have started to incorporate social preferences into individual utility functions. For instance, Andreoni (1990)'s model integrates altruism (or warm-glow), Rabin (1993)'s model integrates a taste for fairness, Fehr & Schmidt (1999)'s model integrates aversion to inequality, while Charness & Rabin (2002)'s model synthesizes previous models by embedding distinct social preferences into a single utility function. Several other papers follow this approach by integrating a variety of social preferences in an individual's utility function. The purported objective of this (more psychologically realistic) endeavor is to try to fit the behavior of experimental subjects: as noted by DellaVigna (2009, p.336, emphasis added), these models have been proposed to "*rationalize* the behavior in these experiments". In a recent review of behavioral game-theoretic modeling, Camerer & Ho (2015, p.560) add that "the hope is that a reasonable utility specification will emerge so that data can be reasonably explained by standard analysis given that new type of utility".

Most of these models do fit experimental or field behavior, but very often in rather limited settings. Rabin (2013, p.622) laments that "once one steps outside of very circumscribed settings, facets of these models clearly correspond to parameter values that fit worse than the baseline self-interested model they aim to replace". Nevertheless, when experimental behavior is inconsistent with a given model, researchers tend to call upon another to rationalize it. If a behavior can not, for

instance, be explained by aversion to unequal outcomes, then one can rely on a different model with a different utility function (Camerer & Ho 2015, p.563). As a result, researchers have at their disposition a collection of models they can use to describe several facets of human social behavior. While this approach is descriptively powerful, allowing researchers to rationalize a large collection of behavior, its predictive power remains very limited.

The central issue is that the existence of social preferences has been inferred *post-hoc* from the behavior of subjects in the laboratory, and different models with different variables and parameters have been constructed to try to make intuitive sense of observed deviations from the predictions of the standard model. The worry is that this *post-hoc* rationalization is not grounded in any theoretical framework. Arguably, without an understanding of why humans might have social preferences and what mechanisms might be giving rise to them, our understanding of human social behavior is unlikely to improve. In fact, Fudenberg (2006, p.699, emphasis added) has already noted that modelers "should devote more effort to *synthesizing existing models* and developing more general ones, and less effort to modeling yet another particular behavioral observation".

What might a general theory of human social behavior (and human social preferences in particular) look like? According to Rabin (2013), useful theories need to have *power* (i.e., they should tell us what not to expect and make sharp predictions that distinguish them from competing theories) and *scope* (i.e., they should apply to a broad set of situations). In this chapter, I argue that the theory of social evolution has the necessary scope and power to provide a useful theoretical framework for human social preferences. Social evolution theory applies to any social interaction and it makes sharp predictions about the kinds of social traits that can be observed in the population. Importantly, it predicts the context-dependent nature of social preferences, given that individuals are expected to respond to the (appropriately defined) benefits and costs of cooperative acts. Since social preferences are context-dependent, social evolution theory can not predict the exact content of social behavior in every situation. Rather, we need to take into account the (social) environment in which individuals find themselves. Therefore, I argue that to better understand the wide variation in the expression of social preferences recorded in the laboratory and in the field, social evolution theory needs to be complemented with insights from cultural (and institutional) evolution and behavioral ecology. Cultural and institutional evolution predict that social preferences will be sensitive to the content of rules that regulate social interactions, while behavioral ecology predicts that social preferences will be sensitive to variables influencing the costs and benefits

of cooperation in different environments. I propose that these two complementary mechanisms can go a long way in helping us predict the variable expression of human social preferences.

2.2 Social evolution and social emotions

According to Fehr & Fischbacher (2002, p.2), "[A] person exhibits social preferences if the person not only cares about the material resources allocated to her but also cares about the material resources allocated to relevant reference agents". The theory of social evolution provides a framework—grounded in the theory of evolution—for thinking about the logic governing social interactions. It is a useful starting point in that it makes clear the conditions under which we should expect organisms, including humans, to engage in behaviors that have consequences not just for themselves but also for others. As such, it helps to better understand the conditions under which humans would exhibit social preferences. Moreover, such a perspective illuminates the function of the emotions underlying social preferences, such as fairness, gratitude, shame or sympathy. This section is, therefore, divided in two parts. The first part describes the logic of social evolution, while the second part describes how this logic illuminates the function of social emotions.

2.2.1 The logic of social evolution

The observation that every organism appears to be designed to do something leads to two important questions: What exactly is this something (what is the *purpose*)? Who, or What is the designer (what is the underlying *process*)? Darwin's theory of evolution (Darwin 1871) parsimoniously addresses both of these questions. The purpose is to maximize the individual's reproductive success; as such, organisms will appear as adapted (or designed) in order to maximize this quantity. The process by which adaptation occurs is natural selection: genes (and traits) that are associated with greater individual reproductive success are expected to increase in frequency in the population; that is, "natural selection acts to increase the mean fitness of individuals in a population" (West et al. 2011, p.232).

Since Darwin published his pioneering book in 1871, there has been only one fundamental change in our understanding of the process of evolution. This fundamental change has arisen from the work of Hamilton (1964). Darwin had trouble explaining a lot of cooperative—and sometimes even sacrificial—acts observed in nature (e.g., worker bees or ants that never reproduce) through the lens of *individual* fitness maximization. How could the theory of evolution by natural selection

explain all of design, if it led organisms to postpone their own reproduction in order to favor others? This apparent paradox derived from the fact that Darwin did not consider *indirect* fitness benefits. Darwin's theory was based on individual reproductive success: as such, he only considered how an organism could maximize its own direct fitness. What Hamilton (1964) has shown is that if one takes the gene-eye's view—that is, the perspective of a single gene (Dawkins 1976)—then it is possible for a gene to perpetuate itself not only directly (by influencing traits that increase the host's direct fitness) but also indirectly (by influencing traits that lead the host to help other organisms which share the same gene). This is known as *inclusive fitness theory*: the genes (and traits) favored by natural selection will be those that maximize an organism's inclusive fitness (the sum of its direct and indirect fitness benefits). The most obvious case in which an organism can increase its fitness indirectly is by helping a kin.[1] As such, through the lens of inclusive fitness theory, the reproductive sacrifice of a worker bee or ant is not surprising. Bees and ants live in colonies in which relatedness is very high;[2] hence, if workers sacrifice for their sisters or their queen, then their genes are unlikely to be wiped-out of the population. Importantly, inclusive fitness theory is not a substitute to the Darwinian theory of evolution. Rather, it "is our modern interpretation of Darwinian fitness in its most general form, explaining both the process and purpose of adaptation" (West et al. 2011, p.233). It follows from this discussion that organisms should behave *as if* they were maximizing their inclusive fitness (West & Gardner 2013).

With this in mind, we can now discuss social traits from an evolutionary perspective. Social traits affect at least one individual other than the bearer of that trait. A useful typology of social traits is provided by West et al. (2007). Social traits can be *selfish*, *altruistic*, *mutually beneficial*, or *spiteful*. These traits are defined with respect to the benefits and costs they entail for the actor and the recipient. Importantly, in evolutionary biology, benefits and costs are defined on the basis of the lifetime fitness consequences of a behavior (Hamilton 1964). That is, a trait needs to be judged based on its average consequences (relative to other traits) over a lifetime. Therefore, while rational choice theory expects organisms to maximize (subject to constraints) in every real-time situation, evolutionary theory emphasizes that a trait should be judged with respect to its average fitness consequences over a lifetime. This contrast is also at the core of Aumann (2019)'s

[1] This is why inclusive fitness theory is also often termed *kin selection* (Smith 1964). Yet, kin discrimination is only one mechanism by which altruistic traits can evolve (limited dispersal being another potentially important mechanism).

[2] The coefficient of relatedness, noted r, is a statistical construct describing the probability (relative to the population average) that two organisms carry the same copy of a gene (allele). For sister bees, for instance, $r = \frac{3}{4}$. For human siblings, $r = \frac{1}{2}$.

distinction between *act*-rationality (maximization of utility in every situation) and *rule*-rationality (adoption of behavioral rules that are adaptive in commonly occurring situations). Consequently, a trait is expected to be "rational", or adaptive, only in those circumstances in which it was selected for.

Selfish traits are traits that benefit the actor at the expense of the recipient (+/-). These traits are favored by natural selection at the individual level as long as they are not directed towards closely related (high r) individuals. *Altruistic* traits are traits that cost the actor at the benefit of the recipient (-/+). These traits are therefore selected against (at the individual level) by natural selection. Famously, they can evolve only if they respect *Hamilton's rule*. Hamilton's rule states that an altruistic allele will be favorably selected if and only if the following relation holds:

$$br > c, \tag{2.1}$$

where b represents the lifetime fitness benefits for the recipient, r represents the coefficient of relatedness between the actor and the recipient, and c represents the lifetime fitness cost for the actor. That is, altruism can evolve if and only if the cost (c) to the actor is lower than the benefit (b) to the recipient times the degree of relatedness (r). As a consequence, altruistic acts are almost always mediated by kinship (cases in which r is high). *Mutually beneficial* traits are defined as traits that benefit both the actor and the recipient (+/+). These are the traits principally studied by economists, often described as reciprocal traits ("You scratch my back, I scratch yours"; Trivers 1971). *Spiteful* traits are traits that cost both the actor and the recipient (-/-). These traits are unlikely to be common in humans (West & Gardner 2010).

Selfish traits are not a source of surprise for economists, given that humans were supposed to be endowed with only such traits. As such, in this chapter, we will mainly consider *cooperative* traits, which are traits that provide benefits to others; they, therefore, encompass altruistic (-/+) and mutually beneficial (+/+) traits. These traits have concerned economists since they have brought humans into the laboratory. By definition, cooperative acts are not necessarily altruistic (in the biological sense). In fact, according to *Hamilton's rule*, truly altruistic acts are expected to be directed almost exclusively toward close relatives. This is so because they otherwise would not be able to evolve in the population (Dawkins 1979, Hamilton 1964). Therefore, costly behavior towards genetically unrelated individuals is expected to be repaid one way or another (or at least to be incentivized in some way; see Section 2.4) (Kurzban et al. 2015).

These important insights stem from the mathematics underlying the theory of

social evolution. Yet, it is one thing to describe what the logic of evolution leads us to expect but another to describe how organisms might implement strategies that respect this logic. The next section describes how social emotions regulate our behavior in such a way as to implement (evolutionary) successful strategies in the social domain.

2.2.2 The logic of social emotions

As described in the Introduction, when economists have brought humans into the laboratory, they found that they tend to exhibit a wide range of emotions that regulate their interactions with others, such as fairness, guilt, shame, anger, warm-glow or empathy. That is, humans appear to be concerned about the welfare of others (Fehr & Fischbacher 2002). These emotions influence whether individuals will be motivated to help, contribute to public goods, and more generally, cooperate with others, or alternatively, punish, or free-ride.

Evolutionary theory illuminates the function of our social emotions. While emotions are part of our everyday language, defining them remains a challenge. In this chapter, we will use the following definition, which emphasizes their evolutionary roots:

> [Emotions] are sensory-motor integrative devices that serve specific adaptive purposes. They are tuned to detect information relevant to particular kinds of environmental challenges and opportunities, and they use this information to control behavioral responses and internal physiological adjustment that help bring closure to the situation. (LeDoux 2012, p.655)

Hence, emotions are proximate mechanisms regulating our behavior and they respond to sets of cues from the environment. As a matter of example, the emotion of fear responds to cues of immediate danger in the environment and triggers a set of mechanisms that influence the organism's attention, motivation, physiology and behavioral decision rules. Importantly, emotions are tuned to and activated by stimuli which are either innate or learned. As such, emotional mechanisms track rewards and punishments (incentives) in our environment and modulate (approach or withdrawal) instrumental behavior (LeDoux 2012). Yet, instead of treating emotions as context-dependent regulatory mechanisms, economists have often treated them as individual traits, an assumption implicitly made when one, for instance, introduces warm-glow directly into an individual's utility function (suggesting that individuals have a preference for altruism).

We can now connect the logic of social evolution to the logic of social emotions. Social emotions represent "a subset of emotions designed to solve adaptive problems of sociality" (Sznycer & Lukaszewski 2019, p.396). The mathematics of social evolution tell us, for instance, that altruistic behavior (in the biological sense) should be directed only towards close relatives, thereby contributing to the inclusive fitness of the actor. How could natural selection design organisms such that they would implement such a strategy? One way would be to provide organisms with (i) a mechanism for recognizing kin and (ii) a mechanism for motivating individuals to help kin preferentially. There is abundant evidence that such mechanisms exist in humans (Lieberman et al. 2007, Lieberman & Lobel 2012, Sznycer, De Smet, Billingsley & Lieberman 2016) and non-human animals (Krakauer 2005, Sherman 1977). Therefore, cues from the environment (e.g., that one interacts with a kin) trigger mechanisms (emotions, such as affection or kindness) designed to motivate individuals to behave cooperatively. In fact, there is compelling evidence that human interactions with kin tend to be more cooperative (Alvard 2009, Kurland & Gaulin 2005). As such, by understanding the logic of social evolution, one can better understand the often powerful emotions mediating our relationship with kin and predict that they will tend to apply only in restricted circumstances (when interacting with kin).

What about interactions with genetically unrelated individuals, which usually comprise the great majority of our relationships? The theory of social evolution predicts that individuals should engage in cooperative acts only if they are repaid in some way or another (or, more generally, only if they are incentivized to). Again, how could natural selection shape organisms such that they would behave in accordance with such a constraint? It would need to provide organisms with mechanisms that (i) give preference to partners both able and willing to provide benefits, (ii) build and maintain long-lasting relationships with valuable partners and (iii) terminate relationships that are no longer beneficial (Beltran et al. 2022). It is widely believed that the social emotions that have been documented by economists have evolved to manage the challenges of regulating, facilitating and maintaining cooperative relationships (Baumard et al. 2013, Beltran et al. 2022, Sznycer & Lukaszewski 2019). For instance, *compassion* or *empathy* can be seen as motivating individuals to help those in need in order to initiate a potential chain of reciprocal favors (Kurzban et al. 2015, Sznycer et al. 2019); the emotion of *anger* can be seen as motivating individuals to bargain for a better treatment, incentivizing others not to take advantage of the actor (Sell et al. 2017); the emotion of *gratitude* can be seen as motivating individuals to reciprocate favors, thereby sustaining relationships (Smith

et al. 2017); the emotion of *guilt* can be seen as motivating individuals to amend wrongdoings and repair existing valuable relationships (Hopfensitz & Reuben 2009, Sznycer & Lukaszewski 2019); finally, the emotion of *shame* can be seen as incentivizing individuals to restore their public image (Sznycer, Tooby, Cosmides, Porat, Shalvi & Halperin 2016).

The previous discussion should make clear that social emotions (or social *preferences*) are neither traits nor individual characteristics. Instead, they are endogenous, context-dependent, and function of the cues present in the environment. While the architecture regulating social emotions is universally shared among humans, the exact conditions that trigger these emotions will differ according to the social environment in which individuals find themselves (Henrich 2020, Sznycer & Lukaszewski 2019). This probably explains the wide variation detected in experiments measuring individual social preferences (Gächter et al. 2010, Henrich et al. 2001). In fact, it is not enough to understand the function of social emotions to predict exactly when they will be triggered. A detailed understanding of the social norms (or institutions) regulating interactions within a given society, as well as the characteristics of the local ecology, are also necessary to make sense of the cross-cultural variation in human cooperative behavior, to which we will return in Section 2.4.

2.3 Is human social behavior outside the scope of inclusive fitness theory?

Economists have argued that human social behavior (and social preferences in particular) can not be explained by standard evolutionary theory. It has been said that "there is more at work in sustaining human cooperation than is suggested by [traditional evolutionary] theories" (Fehr & Fischbacher 2002, p.139), that "current gene-based evolutionary theories cannot explain important patterns of human altruism" (Fehr & Fischbacher 2003, p.785), or that human pro-social behavior is "fundamentally incompatible with the economist's model of the self-interested actor and the biologists' model of the self-regarding reciprocal altruist" (Gintis et al. 2003, p.169). This group of researchers has developed their own theory of human cooperation and altruism, named *strong reciprocity*. Gintis (2000, p.169) writes that "[a] strong reciprocator is predisposed to cooperate with others and punish non-cooperators, even when this behavior cannot be justified in terms of self-interest, extended kinship, or reciprocal altruism". Strong reciprocators therefore cooperate and punish "even though as a result they receive lower payoffs than other group members" (Bowles & Gintis 2004, p.17). While it appears that natural selection should not favor such

types, theoretical models have been developed that support such results (Boyd et al. 2003, Bowles & Gintis 2004, Gintis 2000). Additionally, this group of researchers argues that intergroup competition in our evolutionary past has shaped the evolution of cooperation in humans, such that a process of *group selection*—distinct from individual-level selection—explains the distinctive characteristics of human altruism (Bowles 2006, Bowles & Gintis 2004, Boyd et al. 2003, Gintis 2000). As noted by Burnham & Johnson (2005, p.114), this group of researchers "advocate[s] a 'genuine' altruistic force [...] in human cooperation", one that is not compatible with inclusive fitness theory, in which cooperative acts must necessarily be repaid one way or another (be it directly or indirectly).

So, why does this group of researchers depart from traditional evolutionary theory and what makes them believe that the evolution of cooperation and altruism in humans might not follow the same rules and logic as it does for other organisms? The motivation for their theory stems from observed human behavior in the laboratory (Fehr & Fischbacher 2003). As described in the Introduction of this chapter, when economists have started to test the predictions of traditional theories in economics and game theory, they have discovered that humans are not self-interested and do not play according to traditional equilibrium predictions. Rather, humans take the welfare of (genetically dissimilar) others into consideration, *even in conditions of complete anonymity and without any prospect of future encounter*. Human cooperation can be explained when there are prospects for repeated encounters (Axelrod & Hamilton 1981) or when reputation is at stake (Nowak & Sigmund 1998). But even when such incentives are removed in the laboratory, subjects still contribute to public goods, punish altruistically (Fehr & Gächter 2002) or cooperate even when such cooperation can not be rationalized with self-interest. How, then, can human social behavior be explained with traditional evolutionary theories, if humans behave altruistically towards individuals they are not genetically related to, that they will never meet again, and that can not influence their reputation? These observations, coupled with the mathematical models alluded to above, have led this group of researchers to argue that an alternative to inclusive fitness theory was needed to explain human cooperation and altruism (and social preferences more generally).

The remainder of this section will be dedicated at arguing that human social behavior falls well within the scope of inclusive fitness theory and that human social behavior in the laboratory does not require a thorough rethinking of social evolutionary theory.

2.3.1 Strong reciprocity, group selection and inclusive fitness theory

As described by West et al. (2007) and West et al. (2011), mathematical models on the evolution of strong reciprocity do not show that genuine altruism can emerge; rather, in these models, cooperation provides direct benefits to individuals (through the increased survival of the group to which they belong—a phenomenon known as *group augmentation* (Kokko et al. 2001)) as well as indirect benefits (because these models assume limited dispersal and, therefore, an increased r among individuals inside a group). The discrepancy between what the models show and the conclusions this group of researchers draws from them stems from the particular definition of altruism that they use. In fact, their models show that a "weak" version of altruism can evolve. "Weak" altruism is a behavior that is individually costly within the social group but benefits all group members (including the actor). Therefore, this "weak" altruistic act can provide direct benefits to the actor (through increased group productivity or reduced group extinction) and can evolve even in a population of self-interested individuals. Since the "weak altruistic" act provides direct benefits to the actor, the behavior is considered *mutually beneficial*—not altruistic—using inclusive fitness theory terminology. Hence, contrary to what has been argued, strong reciprocity models are not incompatible with inclusive fitness theory but are well within its scope (Kay et al. 2020).

Similarly, claims that group (or multi-level) selection can explain human altruistic acts that inclusive fitness theory can not predict appear unfounded, since group selection and inclusive fitness theory are mathematically equivalent frameworks (Birch & Okasha 2015, Marshall 2011). Both approaches are just different perspectives on the same underlying process. To see this, we can start with the *Price Equation* (Price 1970), which is a general method for describing evolutionary change from generation to generation. Assume, for simplicity, a population of haploid individuals interacting inside groups of the same size. Let w_i represent individual fitness (or total number of surviving offspring in the next generation) and z_i represent individual i's "genic value" for the altruistic trait, with $z_i = 1$ for altruistic individuals and $z_i = 0$ for selfish individuals. The Price Equation, assuming unbiased transmission of traits, can then be written:

$$\bar{w}\Delta\bar{z} = Cov(w_i, z_i), \tag{2.2}$$

where $\bar{w}$ represents the average fitness of individuals in the population, $\Delta\bar{z}$ describes the average change of the trait from one generation to the next, and $Cov(w_i, z_i)$

the covariance between the trait and individual fitness. The Price Equation tells us that the average frequency of the altruistic trait will increase in the population (from one generation to the next) as long as the covariance between the altruistic trait and individual fitness is positive. The main difference between the inclusive fitness and the multi-level selection approaches is that they decompose the covariance term differently. The multi-level selection approach decomposes the covariance term into *between-group* and *within-group* selection components, such that the Price Equation can be rewritten:

$$\bar{w}\Delta\bar{z} = Cov(w_i, z_i) = Cov(W_k, Z_k) + E_k[Cov(w_{jk}, z_{jk})], \tag{2.3}$$

where W_k represents the average fitness of the kth group, Z_k represents the average altruistic trait value of the kth group, w_{jk} represents the individual fitness of the jth individual in the kth group, z_{jk} represents the "genic" value of the jth individual in the kth group and E_k represents the expectation. Therefore, according to this approach:

$$\Delta\bar{z} > 0 \iff Cov(W_k, Z_k) > -E_k[Cov(w_{jk}, z_{jk})]. \tag{2.4}$$

Since altruism is selected against inside groups, multi-level selection predicts that altruism can evolve ($\Delta\bar{z} > 0$) only if between-group selection (for the trait) is stronger than within-group selection (against the trait). The inclusive fitness approach, on the other hand, rewrites the individual fitness w_i as a regression equation,[3] such that the Price Equation can be rewritten:

$$\bar{w}\Delta\bar{z} = (-c)Var(z_i) + (rb)Var(z_i), \tag{2.5}$$

where $-c$ represents the lifetime fitness costs of the altruistic act, r represents the coefficient of relatedness between partners, b represents the lifetime fitness benefits for the recipient of the altruistic act and $Var(z_i)$ represents the variance of the trait in the population. The first term is the direct effect of the altruistic act, while the second term is the indirect effect of the altruistic act. Therefore, according to the inclusive fitness approach:

$$\Delta\bar{z} > 0 \iff rb > c, \tag{2.6}$$

which is *Hamilton's Rule*. This condition states that for altruism to evolve in the population, it must preferentially be directed towards relatives (see Section 2.2.1).

[3]For details, see Birch & Okasha (2015), Lehtonen (2016), Smith (2020).

To see why the two approaches are equivalent, note that between-group selection (for the trait) is stronger than within-group selection (against the trait) when the within-group disadvantage is suppressed and all the selection happens at the between-group level. This happens only if the variance in the trait is low inside groups and high between groups; in other words, it happens only if individuals within groups are related and highly likely to share the same trait. Therefore, what ultimately matters for the evolution of altruism is positive assortment (altruists preferentially interacting with each other) (Birch & Okasha 2015), and both inclusive fitness theory and multi-level selection capture that insight.

2.3.2 How, then, can we explain human social behavior in the laboratory?

If, after all, human social behavior is compatible with the logic of inclusive fitness theory, how can we explain the apparent systematic deviations from its predictions in laboratory experiments? There are several ways one can explain behavior in experimental settings without invoking the need to rethink social evolutionary theory. I will review two arguments for why human behavior in the laboratory departs from standard evolutionary predictions. Importantly, rather than being mutually exclusive, these arguments are complementary.

2.3.2.A Confusion

Subjects might be confused by the experimental environment and/or the rules of the game and might cooperate more than expected as a result of playing imperfectly. For instance, it has been shown that when full cooperation in public goods games is the dominant strategy, subjects do not contribute 100% of their endowment, suggesting that they either did not understand the game or that they did not place positive value on the welfare of others (Burton-Chellew & West 2013, Kümmerli et al. 2010); alternatively, when no cooperation is the dominant strategy, subjects contribute positive amounts (Burton-Chellew & West 2013). Moreover, when subjects are given information about how their cooperation positively affects other members, they cooperate less compared to a situation in which they do not know their choices affect others (Burton-Chellew & West 2013, Burton-Chellew et al. 2015). This is difficult to reconcile with the idea that individuals have an intrinsic disposition to be fair or care about the welfare of others. In fact, subject types (whether they are categorized as free-riders or fair cooperators) have been shown to depend on the level of understanding of the game, not on an intrinsic predisposition (Burton-Chellew

et al. 2016). These results suggest that higher-than-expected cooperation in the laboratory might result from the imperfect behavior of subjects trying to maximize their own payoffs (Burton-Chellew & West 2021, McAuliffe et al. 2019).

2.3.2.B Spillovers

Evolutionary theory predicts that behavior will be adapted to the environment in which it was selected for. Since humans spend most of their time outside the laboratory, they might bring inside the laboratory patterns of behavior that are adapted to their everyday lives. According to this hypothesis, individuals learn what behaviors are optimal in their social environment and these behaviors *spill over* when they find themselves in unusual and contrived environments. Given that cooperation usually pays off in everyday life and that reputation is generally at stake, humans might develop intuitions which are then brought inside the laboratory and which might explain higher-than-expected cooperation (Aumann 2019, Rand et al. 2012, Rand, Peysakhovich, Kraft-Todd, Newman, Wurzbacher, Nowak & Greene 2014). In fact, cooperation is intuitive only for those individuals who have cooperative relationships outside the laboratory (Rand et al. 2012, Rand, Peysakhovich, Kraft-Todd, Newman, Wurzbacher, Nowak & Greene 2014).[4] Moreover, individuals experimentally exposed to environments supportive of cooperation are more likely to become intuitive cooperators in subsequent one-shot interactions (Peysakhovich & Rand 2016, Stagnaro et al. 2017). As a matter of fact, the spillover hypothesis predicts cross-cultural variations in behavior inside the laboratory (Gächter et al. 2010, Henrich et al. 2001): depending on the content of norms outside the laboratory (are there strong or weak norms of civic cooperation?), individuals in experimental settings will tend to behave more or less cooperatively. Finally, if spillovers from everyday interactions influence behavior in the laboratory, then we can expect individuals to update their behavior once they get familiar with the experimental settings. In fact, there is considerable evidence that subjects used to playing games in laboratory settings converge over time to the optimal strategy (Burton-Chellew & West 2021, Conte et al. 2019, McAuliffe et al. 2018, Rand, Peysakhovich, Kraft-Todd, Newman, Wurzbacher, Nowak & Greene 2014).

Therefore, there are a variety of reasons why humans in laboratory settings might cooperate more than can be expected from the predictions of standard evolu-

[4] While one might retort that non-cooperative *types* might be more likely to have non-cooperative relationships, it has been shown that when intuitive cooperators have time to deliberate (i.e., have time to think about the novel settings), cooperation is reduced (Rand et al. 2012, Rand & Kraft-Todd 2014).

tionary models.[5] Accumulating evidence supports these alternative views, running counter the idea that a radical rethinking of social evolutionary theory is necessary to understand human social behavior.

2.4 How can we explain the variability in the expression of social preferences?

Inclusive fitness theory is a theory of genetic evolution, providing a framework for what kind of traits can be expected in the population. We have seen that if human social behavior respects the logic of inclusive fitness theory, then costly cooperative acts need to (probabilistically) be recovered in the long run, be it directly or indirectly. Yet, while inclusive fitness theory constitutes a powerful organizing framework for thinking about human social behavior, the actual content of social behavior can not be specified *à priori*. In fact, the wide variation in human social behavior within and across societies speaks volumes regarding the importance of the particular social environment in which individuals find themselves (Nettle 2015, Henrich 2020). This section will discuss how ideas from cultural (and institutional) evolution and behavioral ecology need to be integrated to get a fuller understanding of the variability in human social behavior.

2.4.1 Variability in social norms and institutions

A characteristic of humans is that they cooperate extensively in large groups of unrelated individuals. While reciprocity can sustain cooperation at small scales, it is unlikely to underlie the large-scale cooperation that is characteristic of contemporary societies. This is so because reciprocity is unlikely to evolve when group size increases (Boyd & Richerson 1988) and because reciprocal strategies in large groups can lead to universal defection if errors or mistakes are common (Boyd 2017). Rather, strategies involving punishment (or withdrawal of benefits) have been shown to be evolutionarily stable and can sustain cooperation in sizable groups (Boyd & Richerson 1992, Panchanathan & Boyd 2004). The advantage of punishment is that it can get directed towards the defector without punishing other cooperators. For such strategies to be effective, reputations need to flow freely among group members, so that individuals can condition their behavior on the reputation of their partner. This suggests that individuals will be particularly concerned with their

[5] For a discussion on other potential mechanisms leading to cooperation in one-shot interactions, see Raihani & Bshary (2015).

own and others' reputation (Fehr 2004). Yet, these models do not specify what counts as cooperation, nor what counts as defection. To better understand cross-cultural variation in cooperative behavior, it is a useful starting point to consider the content of norms (or institutions) that regulate social interactions.

The emergence and evolution of institutions are thought to have shaped the last major evolutionary transition from small-scale societies characterized by kinship and personal exchange to large-scale societies characterized mainly by impersonal exchange (Powers et al. 2016). Institutions are packages of rules, some of which bear exclusively on the regulation of social life:

> These rules are necessarily recognised and followed by many individuals, and violations are enforced by coordinated sanctioning. [Institutions] define what is normative, and they change the rules of the social game by changing the mapping between individual strategies and the corresponding outcomes, i.e. the payoff matrix. (Powers et al. 2016, p.5)

Social norms can be defined as informal institutions that establish normative behavioral standards (Chudek & Henrich 2011). The norms shared inside a community therefore specify what the theoretical models do not: they describe what counts as cooperation, what counts as defection and what kind of punishment is appropriate. Importantly, in this framework, following norms and rules is not altruistic. As noted by Boyd (2017, p.188), "[n]orms are maintained by rewards and punishments that make it beneficial to follow the norms". Therefore, individuals cooperate (e.g., contribute to public goods) because they are incentivized to, thereby avoiding punishment, shunning or ostracization, potentially even accumulating reputational benefits.[6] Hence, to understand cooperation in large groups, it is important to understand the rules that govern such interactions; i.e., it is important to understand the nature of social norms and institutions governing social interactions.

There is considerable evidence that social norms influence human social behavior, already early in development (Rakoczy & Schmidt 2013, Schmidt & Tomasello 2012). Cross-cultural studies have shown that from middle childhood, children start to become sensitive to the specific local norms surrounding costly cooperative behavior, thereby driving and sustaining the observed cross-cultural diversity in human social behavior (House et al. 2013, 2020). Some researchers have argued that, given the essential role social norms have played in sustaining group cooperation throughout our species' evolution, humans are endowed with a *norm psychology* which facil-

[6] This framework therefore fits within the logic of inclusive fitness theory. Recall that the theory predicts that costly cooperative acts need to be incentivized. The avoidance of punishment can be a strong incentive to cooperate, while reputational benefits can be recovered in the long run.

itates their acquisition (Chudek & Henrich 2011). Such a psychology would explain the early (universal) ontogeny of social norms and would facilitate their acquisition and internalization, thereby potentially explaining why learned rules of behavior spillover when individuals are studied in the laboratory.

If the costly cooperative acts—such as contributions to public goods—that we observe in everyday life mainly result from an adherence to local social norms (and not from altruism towards strangers), then several predictions follow: individuals would want to avoid contributing if possible; individuals would not be sensitive to the impact of their contribution; individuals would condition their contribution on that of others; cooperation should increase when observability is greater. Several studies have confirmed these predictions. For one thing, individuals seem to "avoid the ask", that is, avoid situations in which they might be asked to contribute, either in real-life (Andreoni et al. 2017, DellaVigna et al. 2012, Schwartz et al. 2019) or in the laboratory (Dana et al. 2006). Moreover, they tend to systematically make use of features in the design of experiments to find excuses for not contributing (Exley 2016, Exley & Kessler 2019). Individuals also appear to be insensitive to the impact of their contribution (Burum et al. 2020, Karlan & List 2007, Null 2011). Further, individuals tend to contribute only when it is made clear that contribution is expected from them (Ayres et al. 2013, Goldstein et al. 2008). Finally, when observability is increased and individuals can, as a result, reap the reputational benefits of contributing (or avoid being punished for not contributing), contributions tend to increase significantly (Alpizar et al. 2008, Funk 2010, Rogers et al. 2016, Yoeli et al. 2013). These findings are difficult to reconcile with the idea that humans have an intrinsic predisposition to act altruistically. Instead, it seems that individuals strategically adapt their cooperative behavior to reap reputational benefits or avoid coordinated punishment, which is in line with the theoretical framework outlined in Section 2.2, since cooperation provides direct benefits.

If differences in social norms lead to differences in cooperative behavior, then an important avenue for future research is to understand how social norms and institutions evolve and how to better harness our concern for reputation in order to meaningfully create environments in which cooperation can be individually advantageous. Some have argued that prosocial norms evolve through a process of cultural group selection (Richerson et al. 2016), while others argue that norms and rules are the end-result of a bargaining process among competing parties (Powers et al. 2016, Singh et al. 2017). In any case, a deeper understanding of the interaction between individual psychology and social institutions is necessary in order to improve our understanding of human social behavior (Henrich 2015, 2020).

2.4.2 Variability in the local ecology

Differences in the expression of social preferences can also arise due to differences in the local ecology, that is, in the immediate physical and social environment in which individuals find themselves. While social norms tend to be culturally transmitted, the immediate environment can also be a source of variation in behavior without anyone explicitly or intentionally transmitting information. The idea, which defines the field of behavioral ecology (Nettle et al. 2013), is that individuals are endowed with conditional behavioral rules that have evolved to be sensitive to environmental variations (also known as *adaptive phenotypic plasticity*). The logic is detailed in the following quote:

> *If* in different environments some behaviors are more biologically adaptive than others, *and* organisms have regularly encountered varying environments (across time or location) in their ancestral history, *then* natural selection should favor the evolution of environmentally sensitive flexibilities. (Sng et al. 2018, p.715, emphasis in original)

In this framework, individuals with similar genotypes might nevertheless behave differently if confronted with different environmental cues, such as population density, resource availability and unpredictability, extrinsic mortality, etc. (Nettle 2015, Sng et al. 2018). The behavioral rules underlying behavioral variation need not be implemented consciously: all that is necessary is that the organism adapts its behavior to the cues present in the immediate environment.

The relevance of such a perspective is nicely illustrated by Nettle (2015). In a quantitative ethnography of two similar neighborhoods in the same city, Nettle and co-authors have provided convincing evidence that the particular environment in which individuals find themselves causally influences the expression of so-called social preferences. In their study, one neighborhood is characterized by historical economic deprivation while the other can be described as relatively wealthy. It has been shown, through a series of surveys and experiments, that material or social cues from the environment directly influence trust levels, willingness to give in dictator games or willingness to punish third parties (Nettle et al. 2011, 2014, Schroeder et al. 2014). Importantly, this work shows that what can be described as norms of cheating may actually stem from feedback mechanisms responding to initial isolated events.[7]

[7]For instance, if one observes littering, theft or violence, one might automatically infer that the environment is not safe, therefore influencing one's level of social trust or willingness to cooperate and punish, ultimately influencing the inferences and subsequent behavior of others, and so on, creating a downward spiral that might lead to a stable equilibrium.

The powerful effects of such subtle cues have also been documented by Keizer et al. (2008): by experimentally inducing disorder in a natural environment (e.g., by introducing graffiti), they have shown that people subsequently modify their behavior in accordance (e.g., by littering more or breaking commonly held social rules). Further evidence that individuals calibrate their social behavior to the local ecology comes from Lamba & Mace (2011). By studying *subgroups* from one small-scale society, they show that there exist variations in cooperative behavior that can be traced to demographic and ecological variations between groups. This work provides conclusive evidence that individuals regulate their social behavior as a function of the cues present in their environment, confirming the idea that viewing humans as having stable prosocial dispositions (or preferences) is not supported by the evidence. This work also creates a link between material/ecological conditions and social norms, thereby potentially casting a light on the process of norm emergence and the particular form they take (Sng et al. 2018).

Key to the behavioral ecology approach to human behavior is the idea of *internal regulatory variables* (Tooby et al. 2008). As described by Nettle (2015, p.99), "[i]nternal regulatory variables are running mental meters of some aspect of the environment or your own state". The idea is that these variables are being fed with informational inputs coming from the environment and adjusted as a consequence. For instance, an individual's social trust level can be seen through this lens: if there are enough cues in the local ecology that others can be trusted, then the variable might be calibrated upwards (and conversely). Analogously, one might decide to behave cooperatively only if there are enough cues that one's cooperation might be rewarded. This perspective, therefore, emphasizes the context-dependent nature of human social behavior as well as the need to integrate individual studies into their broader environmental and social context. Conclusive evidence for the existence of such regulatory variables comes from cross-cultural studies in the expression of emotions (Sznycer et al. 2017, 2018, Sznycer & Lukaszewski 2019). This group of researchers has shown that human emotions such as shame, pride or gratitude are universal, yet their specific expression depends on the values shared inside different communities. For instance, Sznycer & Lukaszewski (2019) have shown that the greater the social valuation for a given trait or act, the greater the associated elicitation of a social emotion associated with that trait or act (e.g., if in culture A trustworthiness is more valued than in culture B, then engaging in an untrustworthy act in culture A leads to a greater elicitation of shame—to avoid devaluation by others—than it does in culture B). This suggests that the expression of social emotions mediating human social behavior is responsive to cues in the local environ-

ment, including the content of shared norms of behavior, allowing humans to adapt to and successfully navigate their social world.

2.5 Theoretical application: understanding *moral wiggle room*

This section will be dedicated at illustrating the ideas developed in this chapter with the help of a simple theoretical model. The objective will be to link a quirky feature of social preferences to benefits and costs (incentives) in the social environment.

More specifically, the model will apply to what has been termed "moral wiggle room" (Dana et al. 2007). Moral wiggle room describes the individual tendency to exploit features of the environment in order to avoid behaving prosocially. More precisely, when researchers introduce noise in their experimental settings, rendering the relationship between individual actions and outcomes less transparent, subjects tend to take advantage of this, reducing their contributions compared to a baseline in which ambiguity is removed (Dana et al. 2007, Di Tella et al. 2015, Grossman & Van der Weele 2017). That is, individuals tend to take advantage of situational ambiguities to behave selfishly while simultaneously maintaining enough plausibility to avoid being seen as selfish. These results have challenged the idea that individuals have stable prosocial dispositions, since individuals endowed with other-regarding preferences are not expected to take advantage of features of the experimental design to avoid donating to others. Nevertheless, researchers have not yet tied "moral wiggle room" to incentives in the individual's social environment. Rather, researchers usually assume that the tendency to take advantage of situational ambiguities to avoid behaving prosocially reflects a preference for a positive self-image or self-signaling (Grossman & Van der Weele 2017).

I want to describe here, based on work by Panchanathan & Boyd (2004) and Hoffman et al. (2018), that the tendency to take advantage of situational ambiguities can be expected if prosocial behavior is socially enforced and enforcement is coordinated (see Section 2.4.1). Panchanathan & Boyd (2004) have shown that prosocial behavior (modeled as a contribution to a public good) can be sustained in settings in which contributions are socially incentivized, that is, in settings in which non-contributions are punished by third-parties upholding a norm. In their model, individuals behave prosocially by fear of being in bad standing (and therefore, having help from others withdrawn) and punishment of non-contributions is incentivized by the fear of higher-order punishment. Punishment is therefore coordinated, with each player punishing non-contributions because they expect others to punish them

if they do not. While Panchanathan & Boyd (2004) introduce potential errors in their model (e.g., a cooperator might mistakenly fail to help a contributor in good standing), they do not consider situations in which it is unclear whether individuals have contributed or not. That is, contributions and non-contributions are publicly observed and common knowledge. This assumption is crucial, since punishing players are expected to exclusively punish non-contributions, at the risk of losing good standing if they punish a player who has contributed. One might ask, what are the predictions of the model in situations in which player contributions are not common knowledge?

Hoffman et al. (2018) have investigated the conditions under which coordinated punishment can be stable (can be a Nash equilibrium). They model coordinated punishment as a coordination game between two players. Both players would want to punish if and only if they expect the other player to be sufficiently likely to punish himself (formally, the other player needs to punish with probability greater than $\bar{p}$, with $\bar{p}$ the risk-dominance of the coordination game). Importantly, before deciding whether to punish or not, both players receive signals on which they can condition their decision. Hoffman et al. (2018) formally describe the signal structure that allows punishment to be triggered. They show that for players to be motivated to punish, the signals need to be dually $\bar{p}$-evident. In other words, punishment can be incentive-compatible if signals are sufficiently correlated and/or observable. As a result, public signals are particularly effective at triggering sanctions, while private, uncorrelated signals tend to prevent coordination (and therefore punishment).

The following model merges these two approaches so as to investigate the conditions under which contributions to a public good can be stable, when contributions might not be common knowledge among the players.

2.5.1 Outline of the model

The game is played between one agent and two enforcers (Enforcer 1 and Enforcer 2):

1. The agent takes an action $a \in \{C, \bar{C}\}$, with C being interpreted as *contributing* (or *behaving prosocially)*, and $\bar{C}$ being interpreted as *not contributing* (or *behaving selfishly*). Playing C costs the agent an amount $c > 0$, while playing $\bar{C}$ does not cost anything.

2. The agent's decision influences the state of the world. Let $\Omega = \{C, \bar{C}\}$, with the state of the world C being interpreted as *the agent has contributed*, and state of the world $\bar{C}$ being interpreted as *the agent has not contributed*. The

agent knows the state of the world. Enforcers have common prior beliefs $\mu : \Omega \to \mathbb{R}$ over the set of states of the world, with $\mu(C) = \mu_C$ and $\mu(\bar{C}) = \mu_{\bar{C}}$, and $\mu_C + \mu_{\bar{C}} = 1$.

3. The enforcers have to infer the state of the world $\omega \in \Omega$ from signals. The set of possible signals is $S = \{C, \bar{C}\}$. Receiving signal C has to be interpreted as the enforcer believing state C has realized, while receiving signal $\bar{C}$ has to be interpreted as the enforcer believing state $\bar{C}$ has realized. It is assumed that there are no *false negatives*: when $\omega = C$, enforcers can not receive a signal of $\bar{C}$. That is, if the agent has contributed (played C), then enforcers can not possibly believe the agent has not contributed. The interpretation is that contributions are readily and publicly observed. Therefore, it is assumed that $P(s_i = C|\omega = C) = 1$, with $i = \{1, 2\}$. However, there might be *false positives*: when $\omega = \bar{C}$, enforcers can potentially receive a signal of C. That is, when the agent has not contributed, enforcers might not be certain that she has not. The interpretation is that even if the agent has not contributed, she might plausibly argue—and convince enforcers—that she actually has, or that she did not in fact had the opportunity to contribute. It is therefore assumed that $P(s_i = \bar{C}|\omega = \bar{C}) = 1 - \epsilon$, with $i = \{1, 2\}$, which can be interpreted as the *observability* (or *verifiability*) of the state of the world $\bar{C}$. The frequency of false negatives is therefore ϵ, and we write $P(s_i = C|\omega = \bar{C}) = \epsilon$, with $i = \{1, 2\}$. The smaller ϵ, the greater the state of the world $\bar{C}$ can be said to be observable and the less likely it is for the agent to be able plausibly defend her non-contribution; conversely, the greater ϵ, the less observable the state of the world $\bar{C}$ is and the more likely the agent can plausibly defend her non-contribution. Finally, we assume that signals are independently drawn, such that $P(s_i = \bar{C}|s_{-i} = \bar{C}, \omega = \bar{C}) = 1 - \epsilon$, with $i = \{1, 2\}$.

4. Enforcers then play a coordination game. Based on the signals that they have received, they can decide whether to *punish* the agent (play X) or *not punish* the agent (play Y). The interpretation is that both enforcers would like to punish the agent only if they expect the other enforcer to punish too, and conversely (Boyd et al. 2010, Molleman et al. 2019, Panchanathan & Boyd 2004). The payoff matrix of the coordination game is the following:

		Enforcer 2	
		X	Y
Enforcer 1	X	(x, x)	(m, n)
	Y	(n, m)	(y, y)

For it to be a coordination game, we need to set $x > n$ and $y > m$. The *risk-dominance* of the coordination game is written $\bar{p} = \frac{y-m}{(y-m)+(x-n)}$. The interpretation is that for Enforcer 1 to be willing to play X (punish) in the coordination game, she must expect Enforcer 2 to play X (punish) with a probability greater or equal to $\bar{p}$, and conversely. It is assumed that if both enforcers decide to punish the agent (play X), then the agent incurs a cost $z > 0$. If only one enforcer decides to punish the agent, then the agent incurs a cost v, with $0 < v < z$. If both enforcers decide not to punish the agent (play Y), then the agent incurs no cost. Note that the only incentive for Enforcers is to coordinate their decisions, independently of the agent's strategy. Therefore, their payoffs are state-independent.[8]

2.5.2 Equilibrium specification

A strategy for the agent is a choice of action $a \in \{C, \bar{C}\}$. A strategy for the enforcers is a choice of action (X or Y) as a function of the signal received. Therefore, $\sigma_i : S \to \{X, Y\}$ represents enforcer i's strategy, with $i = \{1, 2\}$. For simplicity, we consider only pure strategies.

The main question of interest is the following: under what conditions can the agent be incentivized to contribute (play C)? It is already apparent that the agent will be willing to contribute only if she expects enforcers to punish her if she does not; otherwise, there would be no incentives to pay the costs of contributing c. Therefore, this amounts to asking under what conditions can the strategy profile $\{C, \sigma_1^*(s_1), \sigma_2^*(s_2)\}$, with $\sigma_i^*(s_i) = X$ if and only if $s_i = \bar{C}$, for $i = \{1, 2\}$, be a Bayesian Nash Equilibrium (BNE) of the game.

Proposition 2.5.1. *The strategy profile $\{C, \sigma_1^*(s_1), \sigma_2^*(s_2)\}$, with $\sigma_i^*(s_i) = X$ if and only if $s_i = \bar{C}$, for $i = \{1, 2\}$, is a Bayesian Nash Equilibrium of this game if and only if the following conditions are satisfied:*

[8] This assumption might appear extreme, since we can expect Enforcers to prefer "punish when I believe the agent has not contributed" compared to "punish when I believe the agent has contributed". This admittedly simplifying assumption is made so as to concentrate the analysis on the Enforcers' incentive to coordinate their response.

1. $1 - \epsilon \geq \bar{p}$,

2. $\bar{p} \geq \frac{(1-\mu_C)\epsilon(1-\epsilon)}{\mu_C + \epsilon(1-\mu_C)}$,

3. $c \leq (1-\epsilon)[(1-\epsilon)z + 2\epsilon v]$.

Proof. *In the Appendix.*

Conditions (1) and (2) are analogous to Conditions (5) and (6) from Hoffman et al. (2018)'s Proposition 2, respectively. The principal idea is that enforcers will be willing to play according to their equilibrium strategy (Punish (play X) when receiving $s = \bar{C}$ and refrain from Punishing (play Y) when receiving $s = C$) if and only if signals are not too noisy. Signal noisiness can arise if the state of the world $\omega = \bar{C}$ is not sufficiently observable or verifiable. In such cases, when ϵ is relatively high, coordination between enforcers can be prevented because both enforcers are not sufficiently confident that the other has received the same signal. If coordination is prevented due to the noisiness of signals, then punishment can not be sustained at equilibrium.

Condition (3) is unique to our setting. This condition needs to be satisfied for the agent to be willing to contribute at equilibrium. We can observe that if $\epsilon \to 1$ the condition is difficult to be satisfied. In fact, when $\epsilon \to 1$, we need $c \to 0$ for the condition to remain satisfied, implying that when the observability of the state of the world $\omega = \bar{C}$ is low (when ϵ is high), the cost of contributing must be negligible for the agent to be incentivized to contribute (play C). Conversely, when $\epsilon \to 0$, that is, when the observability of $\omega = \bar{C}$ is high, the condition is more easily satisfied, therefore suggesting that incentivizing contribution from the part of the agent is more easily achieved when observability is high or when non-contribution is easily verifiable by the enforcers.

The model predicts that individuals will be incentivized to contribute (behave prosocially) only when non-contributions are sufficiently *observable* by others, that is, only when there can be no ambiguity about the individuals' decision not to contribute. When ambiguity is high enough, that is, when non-contributions are not sufficiently observable, individuals can plausibly justify their non-contribution or argue that they did not actually have the opportunity to contribute, thereby preventing coordinated punishment by third-parties. This, it is argued, underlies phenomena such as "moral wiggle room" (Dana et al. 2007), where individuals take advantage of situational ambiguities to avoid behaving prosocially.

2.6 Conclusion

This chapter presents an answer to recent calls from behavioral economists and game theorists to synthesize recent findings from the field and the laboratory into an overarching theoretical framework. Specifically, this chapter aims at providing theoretical underpinnings to what have been called *social preferences*. Social preferences have been attributed to humans once it has been discovered that they tend to take the welfare of others into account when making decisions in social dilemmas. Yet, the existence of social preferences has been posited *post-hoc*, by trying to rationalize existing findings that individuals do not play as predicted by standard models. This has proved problematic, giving rise to a wealth of data without any theoretical framework to make sense of it.

In this chapter, I argue that the theory of social evolution constitutes a useful framework for thinking about human social behavior and human social preferences in particular. I believe that it has the necessary *scope* (applying to any type of social interaction), and *power* (clarifying what can and what can not be expected) to unify disparate findings, allow better inferences from existing data and generate useful predictions. The theory of social evolution illuminates the function of our social emotions, which instead of being individual traits (or preferences) are context-dependent, endogenous mechanisms, which have evolved to allow humans to regulate their social relationships with others. The endogeneity of these mechanisms underlies the wide variation in the expression of social preferences documented cross-culturally.

The endogenous nature of social emotions requires us to investigate the cues that individuals use to regulate their social behavior. I have discussed two important factors which have been shown to matter in the expression of social preferences: social norms and social ecology. A wealth of evidence describes how human social behavior is conditional on the cues present in the local environment, be they socially transmitted or not. As such, rather than considering humans as possessing social *preferences*, which have often been described as individual traits which the individual either is endowed with or not, a better route for improving our understanding of human social behavior would be to acknowledge its context-dependent nature and to identify the incentives underlying its expression.

2.7 Appendix

Proof of Proposition 2.5.1. (1) For enforcers to be willing to play X when receiving the signal $s_i = \bar{C}$, they must believe that the other enforcer has also received this signal (i.e., that the other enforcer will also play X) with probability greater than $\bar{p}$. Therefore, we need:

$$\begin{aligned}
P(s_{-i} = \bar{C} | s_i = \bar{C}) &\geq \bar{p} \\
\frac{P(s_{-i} = \bar{C}, s_i = \bar{C})}{P(s_i = \bar{C})} &\geq \bar{p} \\
\frac{(1-\mu_C) P(s_{-i} = \bar{C}, s_i = \bar{C} | \omega = \bar{C})}{(1-\mu_C) P(s_i = \bar{C} | \omega = \bar{C})} &\geq \bar{p} \\
\frac{P(s_{-i} = \bar{C}, s_i = \bar{C} | \omega = \bar{C})}{(1-\epsilon)} &\geq \bar{p} \\
\frac{P(s_{-i} = \bar{C} | \omega = \bar{C}) P(s_i = \bar{C} | s_{-i} = \bar{C}, \omega = \bar{C})}{(1-\epsilon)} &\geq \bar{p} \\
\frac{(1-\epsilon)(1-\epsilon)}{(1-\epsilon)} &\geq \bar{p} \\
1 - \epsilon &\geq \bar{p}.
\end{aligned}$$

(Note that in line 3, we take advantage of the fact that $P(s_{-i} = \bar{C}, s_i = \bar{C} | \omega = C) = 0$ and $P(s_i = \bar{C} | \omega = C) = 0$).

(2) For enforcers to be incentivized to play Y when receiving signal $s_i = C$, they must believe that the other enforcer will play Y with sufficiently high likelihood. More precisely, enforcers need to believe that the other enforcer will play X with probability lower than $\bar{p}$, or, alternatively, that the other enforcer's probability to receive the signal $s_i = \bar{C}$, given that we have received $s_i = C$, is lower than $\bar{p}$. Therefore, we need:

$$\begin{aligned}
P(s_{-i} = \bar{C} | s_i = C) &\leq \bar{p} \\
\frac{P(s_{-i} = \bar{C}, s_i = C)}{P(s_i = C)} &\leq \bar{p} \\
\frac{(1-\mu_C) P(s_{-i} = \bar{C}, s_i = C | \omega = \bar{C})}{\mu_C P(s_i = C | \omega = C) + (1-\mu_C) P(s_i = C | \omega = \bar{C})} &\leq \bar{p} \\
\frac{(1-\mu_C) P(s_{-i} = \bar{C}, s_i = C | \omega = \bar{C})}{\mu_C + (1-\mu_C)\epsilon} &\leq \bar{p} \\
\frac{(1-\mu_C) P(s_{-i} = \bar{C}, s_i = C | \omega = \bar{C})}{\mu_C + (1-\mu_C)\epsilon} &\leq \bar{p}
\end{aligned}$$

$$\begin{aligned}
\frac{(1-\mu_C)[(1-\epsilon) - P(s_{-i} = \bar{C}, s_i = \bar{C}|\omega = \bar{C})]}{\mu_C + (1-\mu_C)\epsilon} &\leq \bar{p} \\
\frac{(1-\mu_C)[(1-\epsilon) - P(s_{-i} = \bar{C}|s_i = \bar{C})(1-\epsilon)]}{\mu_C + (1-\mu_C)\epsilon} &\leq \bar{p} \\
\frac{(1-\mu_C)[(1-\epsilon) - (1-\epsilon)(1-\epsilon)]}{\mu_C + (1-\mu_C)\epsilon} &\leq \bar{p} \\
\frac{(1-\mu_C)[(1-\epsilon) - (1-\epsilon)^2]}{\mu_C + (1-\mu_C)\epsilon} &\leq \bar{p} \\
\frac{(1-\mu_C)\epsilon(1-\epsilon)}{\mu_C + (1-\mu_C)\epsilon} &\leq \bar{p}.
\end{aligned}$$

(Note that at line 6, we take advantage of the fact that $P(s_{-i} = \bar{C}, s_i = \bar{C}|\omega = \bar{C}) + P(s_{-i} = \bar{C}, s_i = C|\omega = \bar{C}) = P(s_{-i} = \bar{C}|\omega = \bar{C}) = 1 - \epsilon$, so that $P(s_{-i} = \bar{C}, s_i = C|\omega = \bar{C}) = (1-\epsilon) - P(s_{-i} = \bar{C}, s_i = \bar{C}|\omega = \bar{C})$).

(3) For the agent to be incentivized to play C given the equilibrium strategy of the enforcers, her expected payoff of playing C ($E(\pi_C)$) must be greater than her expected payoff of deviating to $\bar{C}$ ($E(\pi_{\bar{C}})$). If she plays C, then enforcers observe C with certainty, and her expected payoff is $-c$. If she plays $\bar{C}$, then: (i) with probability $(1-\epsilon)^2$ both enforcers observe $\bar{C}$ and she gets $-z$, (ii) with probability ϵ^2 both enforcers observe C and she gets 0, and (iii) with probability $2(1-\epsilon)\epsilon$, only one enforcer observes $\bar{C}$ and she gets $-v$. Therefore, it must be that:

$$\begin{aligned}
E(\pi_C) &\geq E(\pi_{\bar{C}}) \\
-c &\geq (1-\epsilon)^2(-z) + 2(1-\epsilon)\epsilon(-v) \\
c &\leq (1-\epsilon)^2(z) + 2(1-\epsilon)\epsilon(v) \\
c &\leq (1-\epsilon)[(1-\epsilon)z + 2\epsilon v].
\end{aligned}$$

■

Chapter 3

The Signaling Value of Social Identity in Polarized Environments[1]

Summary

This chapter proposes a theory of social identity adoption and expression, which ties the choice of social identity to material and social benefits present in an individual's social environment. I argue that in an environment in which receivers aim at uncovering the sender's motives and commitments, the beliefs and values adopted by an individual can serve as a signal of trustworthiness. In such an environment, individuals are expected to adopt the social identity which will provide them with the greatest amount of (social) benefits. I formalize this choice in a game-theoretic framework, embedded in a broader niche selection structure. I argue that the main predictions of the model help illuminate several empirical findings, such as the malleability of beliefs and values, the resistance of beliefs and values to evidence, and the existing correlation between beliefs and values and individual-level traits such as personality.

[1]Another version of this chapter has appeared as BETA Working Paper: Wolff, A. (2022), 'The Signaling Value of Social Identity', *BETA Working Paper*, N° 2022-15.

Classification

JEL Classification: C72, C73, D83, D91

Keywords: Beliefs, Social Identity, Social Incentives, Trustworthiness, Values

3.1 Introduction

Social identity—defined as the set of beliefs and values that categorize individuals in some subset (or group) of individuals—can be puzzling. First, social identity can be highly malleable and highly responsive to changes in beliefs and values among members of our social groups and social networks. Second, changing social environments often precede changes in social identity. Third, social identity appears to accommodate positive information but seems resistant to conflicting evidence. Fourth, while there appears to be no *à priori* reason for why this would be the case, social identity is correlated with genes and individual-level traits such as personality.

These four puzzles have not been adequately addressed by existing theoretical work on social identity (Charness & Chen 2020, Shayo 2020). For instance, Akerlof & Kranton (2000) assume that different social groups have different norms about how to behave and that individuals suffer disutility (psychological costs) when they deviate from these prescriptions. But why do individuals suffer psychological costs from deviating from group prescriptions? In another influential paper, Shayo (2009) assumes that individuals derive utility from the social status of their own group but that they also suffer psychological costs from their perceived distance from other group members. Individuals are therefore assumed to seek similarity with other group members, but it is unclear why they would have such a preference. In fact, the existing theoretical work tends to focus on the psychological benefits and costs associated with social identity without necessarily addressing where these benefits and costs come from.

What I propose in this chapter is that (i) viewing social identity as a signal of trustworthiness (or intention to cooperate) and (ii) viewing individuals as adopting the social identity that provides them with the greatest amount of (social) benefits can help explain the puzzles of social identity. So, the argument is that (i) individuals will adopt the beliefs and values that signal their intention to cooperate to others members of their community and (ii) they will choose to cooperate with the community that can bring them the greatest amount of (social) benefits.

The question immediately arises as to how social identity can become a signal of trustworthiness. Building on Loury (1994)'s work, I argue that the choice of a social identity often reveals information about an individual's willingness to cooperate. In an environment in which receivers aim at uncovering the sender's not readily observable motives and commitments and in which different social groups have associated different beliefs, values and ideologies that are in conflict with one another

(i.e., in a *polarized* environment), the beliefs and values adopted by the sender are evaluated against beliefs and values adopted by other senders whose motives and commitments may already be known. In such a context, by adopting specific beliefs and values, senders *pool* (respectively *separate*) from others who adopted similar (respectively dissimilar) beliefs and values and whose motives and commitments are publicly known. The choice of a social identity then essentially signals to others the sender's social commitments. Yet, this tells us nothing about which social identity individuals will decide to adopt. Taking as given the fact that social identity can signal trustworthiness, the theoretical part of this chapter aims at describing which social identity individuals are expected to adopt.

In order to highlight the trade-off in the choice of social identity, I formalize this choice by having a *sender* play two repeated Sequential Prisoner's Dilemma (SPD), each with a specific *receiver* (defined as a group of individuals having adopted a given social identity), with both games being preceded by a signaling stage in which the sender has to choose which social identity to adopt. This strategic interaction is embedded in a broader niche selection structure (Smaldino et al. 2019) so as to add an assortment stage which underlies the trade-off in the choice of social identity. Following Loury (1994)'s insight, receivers condition their strategy in the repeated SPD on the sender's decision in the signaling stage, and they never cooperate with a sender that adopts of conflicting identity. In order to solve for the equilibrium choice of social identity, I assume that (i) the value of the sender's continuation probability across both repeated SPD is a function of the benefits that the sender might reap from cooperating with each receiver and (ii) the sender will decide to cooperate with the receiver that can provide her with the greatest amount of benefits. I show that by adopting a given social identity in the signaling stage, the sender can signal *high continuation probability* in the repeated SPD, therefore reassuring the receiver that she will be around in the future to reciprocate favors.

The theory developed in this chapter makes several predictions which can help explain the puzzles of social identity. First, individuals will be eager to adopt the same social identity as others in their community, so as to appear as trustworthy. This can explain the malleability of social identity, since if the beliefs and values of other members of the community change and you still want to appear on their side, then you are expected to change your beliefs and values too. Second, individuals will trade-off the benefits from cooperating with different audiences when deciding which social identity to adopt. This can explain changes in beliefs and values when individuals join a new community, since new community members likely become the people with which individuals now interact the most and on which they now rely

the most. Third, if social identity signals trustworthiness, then there is no incentive to change it when receiving new information, particularly so if other members of the community do not modify theirs. This can explain the resistance of social identity to conflicting evidence. Finally, in the proposed model, individuals with similar traits join the same social niches. If some beliefs and values are more or less enforced in different social niches—which appears to be the case—then this can explain the existing correlation between social identity and genes and personality.

3.2 The puzzles of social identity

This section describes the four puzzles of social identity that have motivated the present work. The focus will be on political and moral beliefs and values since these are the aspects of social identity for which the best empirical evidence exists.

3.2.1 Social identity can be highly malleable

The first puzzle is that social identity can be highly malleable, meaning that it can be highly responsive to changes in the beliefs and values of other members of our social group or social networks. For instance, in the U.S., when legislators send their constituents a letter that contains their position on salient and controversial issues, constituents are significantly more likely to adopt that same position—even if the position conflicts with their previous position and even if no particular justification is provided by the legislators (Broockman & Butler 2017). Similarly, during Donald Trump's presidency, some Republicans and self-defined conservatives reacted to Trump by just following his opinions—whether those opinions were conservative or liberal (Barber & Pope 2019). So, a cue from their party leader was sufficient to completely shift their policy positions on the minimum wage, tax policy or immigration. This, of course, is not restricted to the U.S. political context. Researchers have found the same result in Denmark, with some Danish citizens completely switching their policy positions on sensible issues when their party switched theirs—even though their new policy position conflicted with their previous one (Slothuus & Bisgaard 2021). Importantly, for the purposes of the argument developed in this chapter, the automatic influence of the stance taken by other members of our social group on our own beliefs and values appears to be strongest in polarized environments (Druckman et al. 2013).

Furthermore, Gould & Klor (2019) have shown, using a long-run panel study, that changes in individual political beliefs and values closely track changes in the core tenets of the political party individuals identify with. Even supposedly stable

aspects of social identity are subject to these influences. For instance, Egan (2020) has shown that Republicans and Democrats shift their ethnic, religious and sexual identities following congruent shifts among other members of their political coalitions. Similarly, Agadjanian & Lacy (2021) show that individual racial identities converge towards the identity enforced in their political party.

3.2.2 Social identity is environment-dependent

The second puzzle is that social identity is environment-dependent, meaning that it seems to be a function of which social identity others in our social environment have adopted. In a longitudinal study that corrects for self-selection into networks, Lazer et al. (2010) have shown that the political views of college students shift over time towards the ones prevalent in their friendship networks. So, the more conservative one's social network is, the more conservative one tends to become over time. Similarly, the more liberal one's social network is, the more liberal one tends to become over time. This is also the case for religious beliefs. Mayrl & Uecker (2011) have shown that the best predictor of changes in religious beliefs among college students is their social network composition—whether or not others are also religious. This, of course, is not confined to college students. Martin & Webster (2020) have shown that when individuals move to new neighborhoods, they similarly tend to adopt the political preferences of their new peers and neighbors. Moreover, in her extensive study of the role of social and political networks in influencing patterns of political behavior, Sinclair (2012) finds that an individual's choice of, e.g., a party identification, is strongly influenced by her immediate social network, which has not been formed based on political preferences.

Laboratory experiments similarly show that individual political values tend to respond to social cues, with individuals often modifying their expressed identities when interacting with others holding dissimilar views, when encountering an ideologically homogeneous audience or when receiving information about peer preferences and values (Connors 2020, Klar 2014, Levitan & Verhulst 2016, Toff & Suhay 2019, Visser & Mirabile 2004).

3.2.3 Social identity can be resistant to conflicting evidence

While individuals are usually eager to update their social identity following changes in beliefs and values among members of their social groups or social networks, they are not so eager—even resistant—to modify their social identity when confronted with conflicting evidence. For instance, Nyhan & Reifler (2010) have shown that

when you provide evidence that their political views are mistaken, individuals often do not take that evidence into account and actively try to counter-argue. In fact, the correction can strengthen their pre-existing views—a phenomenon known as the *backfire effect*. Even more telling are findings that individuals do not reduce their support for a party candidate after learning (and accepting) that they have been told lies (Nyhan et al. 2019, Swire-Thompson et al. 2020).

In their paper on French attitudes towards the Carbon Tax, Douenne & Fabre (2022) show that individuals that are supportive of the reform correctly update their beliefs when receiving information about the benefits of the tax, while those that oppose the tax (predominantly Yellow Vests supporters) tend to discard such information, unless it goes against the tax. Similarly, Schaffner & Roche (2016) take advantage of the salient release of a jobs report—announcing a notable decrease in the unemployment rate—during the 2012 presidential campaign in the U.S. in order to analyze how Democrats and Republicans reacted to the news. They find that Democrats correctly updated their beliefs about the unemployment rate downwards, while Republicans appeared to counter-argue the news, updating their beliefs *upwards*.

3.2.4 Social identity is correlated with genes and personality traits

The last puzzle is that social identity is correlated with genes and individual-level traits such as personality. While there appears to be no *à priori* reason for why beliefs and values might be correlated with genes, research using twin studies has shown that around 40% of the variation in political ideology can be attributed to differences in genetic endowment (Dawes & Weinschenk 2020). Similarly, Gerber et al. (2010) have shown that personality traits are correlated with political attitudes. In particular, Openness to Experience is correlated with liberalism, while Conscientiousness is correlated with conservatism.

3.3 Discussion of the main argument

The argument advanced in this chapter is that (i) viewing social identity as a signal of *trustworthiness* and (ii) viewing individuals as adopting the social identity that provides them with the greatest amount of (social) benefits can help address the four puzzles of social identity listed above. The idea is that individuals will—in polarized environments—strategically adopt the beliefs, ideologies and values (the

social identities) that signal their cooperative intent to others in the communities or networks in which they find themselves. According to the argument presented in this chapter, individual values (or preferences) do not shape the choice of a social identity. Rather, it is the existing incentives in the individual's social environment that influence which social identity she will express and adopt.

The question immediately arises as to how social identity might signal trustworthiness. After all, beliefs and values are internal states that can not be observed by others. My focus in this chapter is on social identity as outwardly expressed and I will therefore consider as a signal the *expression* of one's beliefs and values. Yet, this view need not be inconsistent with the idea that these beliefs and values are internally (deeply) felt (Brewer 1991) since one can expect individuals to internalize those beliefs and values that they are incentivized to hold and express (Melnikoff & Strohminger 2020, Schwardmann et al. 2022).

While the argument in this chapter is that the public expression of one's beliefs and values can signal trustworthiness, this need not necessarily be so. We might actually think of a world in which beliefs and values are completely uncorrelated with intentions to cooperate, where people freely exchange ideas, debate, and argue without making any inferences about their interlocutor's trustworthiness. Yet, this undoubtedly is not an equilibrium of Loury (1994)'s "expression game", which is a game played between senders and receivers. Senders want to persuade (or inform) receivers about the state of the world, while receivers want to form an accurate opinion about that state without being manipulated or deceived by the sender. The problem, of course, is that senders have private information about their motives and commitments, which implies that receivers can not directly observe whether the sender can be trusted to be honest or not. The receivers therefore need to find a way to infer the motives and commitments of the sender from the messages that she sends.

One way to do that is to evaluate the messages sent by senders against messages sent by other senders whose motives and commitments are already known, from historical precedent, for instance. So, if known proponents of the status quo or anti-progressives regularly make a certain argument, then a sender sending a similar message will be categorized as sharing the same motives and commitments—irrespective of her true motives and commitments. If this process of inference is common knowledge among the players, then, by sending specific messages, senders will *pool* with others who sent similar messages (and whose motives and commitments are known) and *separate* from others sending dissimilar messages. This, in fact, constitutes an equilibrium pattern of expression and inference and can be seen

as a kind of social convention governing how senders and receivers are expected to behave.[2] In fact, it appears that we find ourselves at such an equilibrium, since Pietraszewski et al. (2015) have shown that people ascribe specific motives and commitments to others following their agreement or disagreement with politicized statements. Therefore, individuals do not just passively accept that others have certain opinions; they also actively try to infer their motives and commitments from their beliefs and their expressions.

So, the argument is the following: in our (polarized) society, different social groups have associated different beliefs, ideologies and values that are often in conflict with those of other groups, such as when one group favors immigration while the other acts to prevent it, when one group strives to extend rights to disadvantaged minorities while the other favors the status quo, or when one group favors free speech while the other expects speech restrictions on sensitive matters. That is, individuals belonging to different social groups often have worldviews that are at variance, beliefs and values that would lead to different policies and moral views prioritizing different issues (Jacoby 2014, Van Bavel & Pereira 2018). Given the commonly known receiver inference process describes above, a sender who decides to adopt and express the beliefs and values of a given social group will (i) *associate* with others having adopted similar beliefs and values and (ii) *dissociate* from others having adopted different beliefs and values. This choice will have the effect of increasing the trust that others having adopted the same beliefs and values confer to the sender (since the sender has *burned bridges* with other groups and they can therefore expect her to cooperate with them), and decrease the trust that others having adopted different beliefs and values confer to the sender.

So, it is in that sense that beliefs and values (and therefore social identity) can signal trustworthiness. Yet, this tells us nothing about which social identity individuals will decide to adopt in such a polarized environment. The theoretical part of this chapter formalizes the process of social identity adoption—taking as given the fact that social identity can signal trustworthiness—and aims at describing which social identity individuals are expected to adopt.

[2]The reason why truthful expression, coupled with "naive" inference, can not be an equilibrium of this game is that if receivers take messages at face value, then (malevolent) senders are incentivized to deviate so as to manipulate/deceive receivers, and so receivers are similarly incentivized to deviate from their "naive" pattern of inference.

3.4 Game-theoretic analysis of the choice of social identity

This section contains a game-theoretic analysis of the choice of social identity, aimed at describing the main incentives underlying the adoption of social identity. The games (repeated Sequential Prisoner's Dilemmas) are played between a *sender* and two *receivers* (audiences). The games are embedded in a broader niche selection structure that adds an assortment stage that generates the trade-off in the choice of social identity.

3.4.1 Model setup

3.4.1.A Niche selection

Society is characterized by a set N of individuals, a set J of social niches, and a set G of social groups. ***Social groups*** are just collections of individuals that share some core beliefs, values, ideologies and/or norms of conduct. Therefore, throughout the chapter, I will use the term *social group* (or *group*) quite generally to refer to ideologically, socially or culturally defined groups, such as "Liberals", "Conservatives", "Animal Activists", "Christians", "Climate Deniers", "Flat Earthers", etc.

In this chapter, I follow Smaldino et al. (2019) in defining ***social niches*** as particular ways of extracting resources from the environment and/or from other individuals. Importantly, each social niche defines an incentive structure for doing certain things or behaving in certain ways. A typical example is a professional occupation: it allows individuals to extract resources (salary, status, prestige, etc.) from their environment, but different professions have different incentive structures. For instance, if you are a scientist, you can be rewarded for spending all day reading and writing, but not so much if you are an operator or a craftsman. Other examples of social niches include sporting (coach, educator or team player), artistic (band member or independent artist), political (activist, candidate or party member), or social (volunteer) activities, each with its own incentive structure. What is crucial is that different social niches create different payoffs for different personality and cognitive profiles. For instance, a highly introverted person might be more fit to do scientific research than to become a lawyer, a musically gifted person will probably stand out in an orchestra, while a natural analytical thinker might excel at chess. Therefore, different *types* of individuals are more or less at ease in different social niches.

To formalize this idea, every social niche $j \in J$ has an associated *ideal trait*

profile γ_j, that is, an associated vector of (personality and cognitive) traits such that an individual endowed with this exact trait profile would be optimally suited for this niche (Smaldino et al. 2019). Therefore, let $\gamma_j = (\gamma_{j_1}, \gamma_{j_2}, ..., \gamma_{j_p})$, with each γ_{j_p} being a bounded random variable whose value represents niche j's ideal value for the p^{th} trait. The ideal trait profile γ_j that characterizes each social niche j can then be thought of as a description of the incentives faced by individuals in different niches, insofar as individuals will be incentivized to join (respectively depart) niches whose ideal trait profile is similar (respectively dissimilar) to their own trait profile. Each social niche j is populated by a set $N_j \subset N$ of individuals. For simplicity, we assume that all members of a given social niche j belong to social group $g_j \in G$.

In this chapter, we take the perspective of a focal individual, the sender s, who starts the game embedded in a community, subsequently called the sender's *home community* m. Members of m are assumed to belong to the social group g_m, with associated beliefs, ideologies, and values, which have been transmitted to the sender. Therefore, the sender starts the game belonging to the social group $g_m \in G$, hence with a given *social identity* I_{g_m}. The natural interpretation is to consider the sender as having learned the beliefs, ideologies and values (associated with g_m) of the members of the community in which she has grown and developed, with the game being first played while the sender still finds herself in her home community. More generally, this model is expected to apply to any situation in which the sender faces the choice of moving from one (home) community to another, for reasons that are independent of the sender's beliefs and values. Members of the sender's home community constitute the first type of audience (or *receiver*), r_m.

In the first stage, the sender, being embedded in her home community, chooses which *social niche* $j \in J$ she wants to join. The sender s has an associated *trait profile* θ_s, which can be characterized as a vector of P individual behavioral and cognitive characteristics. These together can be said to represent the individual's *cognitive ability* and *personality*.[3] Therefore, let $\theta_s = (\theta_{s_1}, \theta_{s_2}, ..., \theta_{s_p})$, with each θ_{s_p} being a bounded random variable whose value represents individual s's endowment for the p^{th} trait. Importantly, in this first stage, the sender does not choose a social group. This first stage assortment is expected to be a function of the intrinsic characteristics of the individual and of those of the social niche and is, therefore,

[3] An individual's personality represents her traits and behavior that are relatively stable across time and contexts. They are largely innate, in the sense that they are likely "built up from variation in a large number of [...] basal decision-making parameters. Variations in neuromodulatory systems may underlie the differential tuning of these parameters across individuals" (Mitchell 2020, p.124). For an in-depth and insightful discussion about how individuals—in industrialized societies—select and shape their own environments as a function of their innate tendencies, see Mitchell (2020, chapter 5).

not based in any way on individual beliefs or values, which are expected to emerge endogenously from the mechanisms in the model.

3.4.1.B Strategic interaction

The second stage is a *signaling* stage. Once s has decided to join her preferred social niche j, she starts to interact with other individuals (the set $N_j \subset N$) who themselves decided to join that niche (e.g., the sender might decide to join an orchestra which is filled with other music players; she might decide to join a doctoral program in which she interacts with other students; or, alternatively, she might decide to join a law firm in which she interacts with other lawyers). As described above, these other niche members are assumed to belong to the same social group g_j. This simplifying assumption stems from Bonica (2014)'s observation that there exist significant differences in ideological distributions across industries and professional occupations (i.e., social niches), with occupations such as academia, entertainment or media being skewed to the left, while occupations such as banking and finance, building and construction or agriculture being skewed to the right.[4] Other niche members constitute the second type of audience (or receiver), r_j. At this stage, the sender has to decide (i) whether to hold on the beliefs, ideologies and values that she has learned in her home community (i.e., whether to hold on her identity I_{g_m}) or (ii) whether to adopt the beliefs, ideologies and values that prevail among the members of the social niche she has decided to join (i.e., whether to adopt a new identity I_{g_j}). There are, of course, cases in which the social group that is most represented in the home community is the same as in the social niche the sender has decided to join, in which case this model does not bring interesting insights. We will, in this chapter, focus on cases in which I_{g_m} differs from (or conflicts with) I_{g_j} (we can imagine I_{g_m} as representing a Democrat identity, and I_{g_j} as representing a Republican identity).

In the third, *partner choice* stage, the sender s plays a repeated Sequential Prisoner's Dilemma (SPD) with both receivers, r_m and r_j. She therefore plays two games: game m with r_m, and game j with r_j. In the first round of their respective games, receivers decide whether to *Cooperate (C)* in the PD, which implies accepting the sender, or *Defect (D)*, which implies rejecting the sender. Accepting the sender

[4] Audiences (r_m and r_j) are therefore assumed to be homogeneous in terms of beliefs, ideologies and values. In reality, disagreement exists in individual social networks and individuals are therefore unlikely to encounter ideologically homogeneous audiences (Huckfeldt et al. 2013). One way to interpret this assumption is to consider that there is a social group *predominantly* represented among both audiences (receivers) and that the associated beliefs, ideologies and values are enforced by community members. Yet, adding different social groups, more or less equally represented among members of one's community, while potentially closer to reality, is not expected to alter the model's main predictions.

amounts to invest in the relationship with the sender, providing her with a benefit k_i with $i \in \{m, j\}$ at a cost c, with $k_i - c > 0$, while rejecting the sender amounts to refuse to invest in a relationship with her. If the sender is rejected, the game ends, with both players earning payoffs equal to 0. If the sender is accepted, we move on to the second round with probability δ_i, at which point the sender decides whether to *Cooperate (C)* or to *Defect (D)*. Reciprocating (cooperating) also costs c to the sender and provides benefit b to the receiver, with $b - c > 0$, while defecting amounts to bestowing no benefit to the receiver (and paying no cost).

Probability δ_i is meant to capture the idea that the sender might not be there to reciprocate the favor bestowed by the receiver. Hence, δ_i captures the continuation probability of the sender, which is fixed throughout the game.[5] For simplicity, we assume that the sender s can either have a *high (h)* or *low (l)* continuation probability, with $0 < \delta_i^l < \delta_i^h < 1$. Importantly, for our purposes, δ_i need not be the same across both games (i.e., it need not be the case that $\delta_m = \delta_j$). While δ_i can be thought of as exogenously given, it can also be seen as describing the incentives faced by the sender (Jordan et al. 2016, Jordan & Rand 2017): a low continuation probability can realize due to insufficient exposure to mechanisms incentivizing cooperation (e.g., direct or indirect reciprocity, institutions, etc.), while a high continuation probability can stem from high enough exposure to such mechanisms. We will, in this chapter, take this latter perspective, by endogenizing the value of δ_i across both games (see Section 3.4.2.B). Both games are repeated until the relationship between the receiver and the sender is terminated, either because one of the players has defected (played D), or because the sender is not around anymore (i.e., $(1 - \delta_i)$ realizes). That is, we assume that defection from either r_i or s effectively terminates the interaction, reflecting the idea that the other player can not be trusted to cooperate in the future. Finally, if the sender decides not to cooperate with any receiver i, she gets benefit $\bar{\omega} = 0$, which can be considered as the value of her outside option (normalized to zero for simplicity). The structure of the repeated SPD between s and receiver r_i is shown in Figure 3.1. H represents Nature (or chance), and the payoffs are such that r_i's payoff is noted first and s's payoff is noted second.

[5] We assume that the continuation probability of the receivers is 1, such that receivers are always guaranteed to be there for another round.

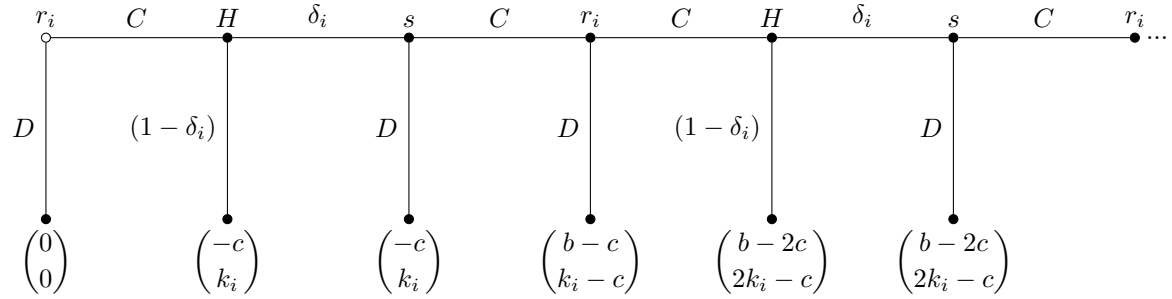


Figure 3.1 – Repeated Sequential Prisoner's Dilemma in the *Partner Choice* stage.

3.4.2 Model resolution

3.4.2.A Partner choice game

We start by seeking Subgame-Perfect Nash Equilibria (SPNE) of the repeated SPD.

Can cooperation (both players playing C throughout the game) be a SPNE of the game? We start by investigating whether it can be beneficial for r_i to play C when s plays C throughout the game. If r_i decides to play C at all of his decision nodes (a strategy we will call $ALLC_{r_i}$) given that s always plays C (plays $ALLC_s$), then his expected payoff $E[ALLC_{r_i}|ALLC_s]$ is:

$$\begin{aligned}
E[ALLC_{r_i}|ALLC_s] &= (1-\delta_i)(-c) + \delta_i(1-\delta_i)(b-2c) + \delta_i^2(1-\delta_i)(2b-3c) \\
&\quad + \delta_i^3(1-\delta_i)(3b-4c) + ... \\
E[ALLC_{r_i}|ALLC_s] &= -c(1-\delta_i)(1+2\delta_i+3{\delta_i}^2+...) + b\delta_i(1-\delta_i)(1+2\delta_i+3{\delta_i}^2+...) \\
E[ALLC_{r_i}|ALLC_s] &= \frac{-c(1-\delta_i)}{(1-\delta_i)^2} + \frac{b\delta_i(1-\delta_i)}{(1-\delta_i)^2} \\
E[ALLC_{r_i}|ALLC_s] &= \frac{(-c+b\delta_i)}{(1-\delta_i)}.
\end{aligned}$$

If r_i decides to play D at the first node, then his payoff will be equal to 0. Therefore, for r_i to be willing to play $ALLC_{r_i}$ when s plays $ALLC_s$, it needs to be the case that:

$$\begin{aligned}
\frac{(-c+b\delta_i)}{(1-\delta_i)} &\geq 0 \\
\delta_i &\geq \frac{c}{b}.
\end{aligned}$$

It follows that as long as $\delta_i \geq \frac{c}{b}$, then r_i is incentivized to play $ALLC_{r_i}$ when s also plays $ALLC_s$. Importantly, if condition $\delta_i \geq \frac{c}{b}$ holds, then r_i is incentivized to play C in every subgame of the game (as long as s also cooperates).

We now investigate the conditions under which s would be willing to play

$ALLC_s$ from her first decision node on, when r_i plays $ALLC_{r_i}$. If s decides to always play $ALLC_s$ when r_i plays $ALLC_{r_i}$, then her expected payoff $E[ALLC_s|ALLC_{r_i}]$ is:

$$E[ALLC_s|ALLC_{r_i}] = (1-\delta_i)(2k_i - c) + \delta_i(1-\delta_i)(3k_i - 2c) + \delta_i^2(1-\delta_i)(4k_i - 3c) + ...$$
$$E[ALLC_s|ALLC_{r_i}] = k_i(1-\delta_i)(1 + \delta_i + \delta_i^2 + ...) + (-c + k_i)(1-\delta_i)(1 + 2\delta_i + 3\delta_i^2 + ...)$$
$$E[ALLC_s|ALLC_{r_i}] = \frac{k_i(1-\delta_i)}{(1-\delta_i)} + \frac{(-c+k_i)(1-\delta_i)}{(1-\delta_i)^2}$$
$$E[ALLC_s|ALLC_{r_i}] = k_i + \frac{(-c+k_i)}{(1-\delta_i)}.$$

If s instead decides to play D at her first decision node, then her payoff will be equal to k_i. For s to be willing to play $ALLC_s$ when r_i plays $ALLC_{r_i}$, then it needs to be the case that:

$$k_i + \frac{(-c+k_i)}{(1-\delta)} \geq k_i$$
$$k_i \geq c.$$

It follows that as long as $k_i \geq c$ is satisfied, then s is willing to play $ALLC_s$ from her first decision node on, as long as r_i plays $ALLC_{r_i}$. As before, if $k_i \geq c$ holds, then s is incentivized to play C in every subgame of the game (as long as r_i also cooperates). We can conclude that as long as $\delta_i \geq \frac{c}{b}$ and $k_i \geq c$ are satisfied, then both players playing C at every decision node is a SPNE of the game. On the contrary, if $\delta_i \geq \frac{c}{b}$ and/or $k_i \geq c$ do not realize, then the only SPNE of the game is for both players to play D at all their decision nodes (a strategy we will call $ALLD$). This result comes from the fact that if the above conditions are not simultaneously satisfied, then at least one player will never play C. But if one player always plays D, then the other is never incentivized to play C, given that they would pay the costs of cooperation c without receiving any future benefits.

Proposition 3.4.1. *There are only two SPNE of the repeated SPD: either (i) both players play D at all their decision nodes or (ii) both players play C at all their decision nodes, when $\delta_i \geq \frac{c}{b}$ and $k_i \geq c$ both realize.*

The proof can be found in Appendix 2. The results of our analysis show that cooperation can be sustained at equilibrium if and only if the receiver is sufficiently confident that the sender will be there in the future to reciprocate the favor (i.e., if and only if δ_i is sufficiently large). Yet, how can the receiver be confident that the sender will be around in the future? Alternatively, how can the sender convince the receiver that she will be around in the future? The key idea is that receivers can

condition their strategy in the partner choice stage to the sender's strategy in the signaling stage. Following our discussion in Section 3.3, we assume that receivers never cooperate with the sender if she has adopted a conflicting social identity. That is, we take as given the fact that social identity can signal trustworthiness and concentrate on which social identity the sender will decide to adopt in such a polarized environment.

3.4.2.B Signaling stage

This section will be dedicated at determining the optimal choice of social identity for the sender. This requires a description of how the values of the sender's continuation probabilities are set across both games.

In the present context, the sender's strategy in the signaling stage ultimately amounts to choose one receiver with whom to cooperate over another, given that her choice of identity I_{g_i} will either attract or alienate some receiver. In particular, if s adopts I_{g_m}, then she knows that only r_m might be willing to cooperate with her. Alternatively, if s adopts I_{g_j}, then she knows that only r_j might be willing to cooperate with her.[6] Hence, if the sender's continuation probability δ_i captures her likelihood of staying around in the future, then the sender's choice of identity in the signaling stage can be seen as analogous to signaling her continuation probability δ_i across both games.

In this chapter, we assume that the sender's continuation probability across both games is a function of the benefits that she might reap from cooperating with r_m and/or r_j. In particular, if the sender is not incentivized to cooperate, then we set $\delta_i = \delta_i^l < \frac{c}{b}$ and cooperation can not stabilize. On the other hand, if the sender is incentivized to cooperate, then we set $\delta_i = \delta_i^h > \frac{c}{b}$ and cooperation can stabilize. Now, what determines whether $\delta_i = \delta_i^l$ or δ_i^h? The answer, it is assumed, lies in the differential benefits that the sender might reap from cooperating with r_m and/or r_j.

To determine the value of the sender's continuation probability δ_i across both games, we start by assuming that $\delta_i = \delta_m^h = \delta_j^h$. That is, we assume that it is equally likely that the sender might be around in both games, reflecting her choice to cooperate with r_m and/or r_j. If $\delta_i = \delta_m^h = \delta_j^h$, we know that the sender is expected to hold on the beliefs, ideologies and values of her home community (hold on I_{g_m}) if the benefits generated from cooperating exclusively with r_m are (i) greater than her benefits from cooperating with r_j and (ii) greater than the value of her outside option. This leads to the following two conditions:

1) $k_m \geq k_j$,

[6] I implicitly assume that receivers can observe the sender's choice across both games.

2) $k_m \geq \delta_i c$.

Condition 1) simply requires that the benefits that she reaps from cooperating with her home community members are greater than the benefits that she reaps from cooperating with other social niche members. Condition 2) is necessarily satisfied at a cooperative equilibrium, which requires $k_m \geq c$. Therefore, if conditions 1) and 2) are satisfied, we set $\delta_m = \delta_m^h > \delta_j = \delta_j^l$, given that the benefits from cooperating exclusively with r_m are greater than the benefits from cooperating with r_j. The sender is then incentivized to cooperate exclusively with r_m, which translates into a higher continuation probability in her game with r_m, which she can signal by holding on the identity I_{g_m}.

Second, by again assuming $\delta_i = \delta_m^h = \delta_j^h$, we know that the sender is expected to adopt the identity I_{g_j} associated with other social niche members if the benefits generated from cooperating exclusively with r_j are (i) greater than her benefits from cooperating with r_m and (ii) greater than the value of her outside option. This leads to the following two conditions:

3) $k_j \geq k_m$,

4) $k_j \geq \delta_i c$.

Condition 3) realizes if the benefit that the sender reaps from cooperating with social niche members is greater than the benefits that she reaps from cooperating with home community members. Condition 4) necessarily realizes at a cooperative equilibrium. Therefore, if conditions 3) and 4) are satisfied, we set $\delta_j = \delta_j^h > \delta_m = \delta_m^l$, given that the benefits from cooperating exclusively with r_j are greater than the benefits from cooperating with r_m. By the same reasoning as before, the greater incentives to cooperate with r_j translate into a higher continuation probability in game j, which she can signal by adopting the identity I_{g_j}.[7]

To summarize, if the benefits from cooperating with r_m are greater than the benefits from cooperating with r_j, the sender s is expected to adopt the social identity I_{g_m} which (truthfully) signals high continuation probability in her repeated interaction with r_m, and therefore can stabilize cooperation between the two players. On the contrary, if the benefits from cooperating with r_j are greater than the benefits from cooperating with r_m, the sender s is expected to adopt the social identity I_{g_j} which similarly (truthfully) signals a high continuation probability in

[7] $\delta_m = \delta_m^l$ and $\delta_j = \delta_j^l$ can not be simultaneously realized since we have assumed that the sender can always decide to cooperate either with r_m or r_j, without constraints. This assumption implies that the gains from cooperating with r_m or r_j can always realize, and these gains are always greater than the gains from defecting. Adopting a social identity I_{g_i} is, therefore, always a truthful signal of high continuation probability in the present setup.

her repeated interaction with r_j. In the present framework, adopting a given social identity therefore signals a high continuation probability in the repeated SPD and can convince receivers to cooperate with the sender.

3.4.2.C Niche selection

In the first stage, the sender s is expected to choose the social niche whose ideal trait profile is the closest (in terms of distance) from her trait profile.[8] This first choice is therefore devoid of any strategic consideration. One can write the (Euclidean) distance d_{sn} between the sender s's trait profile and niche j's ideal trait profile in the following way:

$$d_{sj} = \sqrt{\sum_{p=1}^{P}(\theta_{sp} - \gamma_{jp})^2}$$

Let $\bar{v} : J \to \mathbb{R}_+$ be a function which gives, for each niche $j \in J$, a value $\bar{v}(j) \in \mathbb{R}_+$ to the sender. More specifically, let $\bar{v}(j) = \frac{1}{d_{sj}}$.[9] That is, the smaller the distance between the sender's trait profile and the niche j's ideal trait profile, the greater the benefits the sender can reap from deciding to join j. This value is independent of the benefits that the sender might reap from cooperating with other social members and depends solely on the fit between her trait profile and the niche's ideal trait profile. Given that the distance d_{sj} between the sender's trait profile and the niche's ideal trait profile is fixed and given, the sender is expected, *ex ante*, to choose the niche j^* which satisfies $j^* \in \underset{j \in J}{argmax}\, \bar{v}(j)$.[10]

3.4.3 Equilibrium specification

The following Propositions describe the conditions underlying our two main equilibrium strategy profiles of interest, which characterize the circumstances under which the sender s will be willing to adopt social identity I_{g_m} (Proposition 3.4.2) or I_{g_j} (Proposition 3.4.3).

The sender s is considered to be Player 1, the receiver r_m Player 2, and the

[8] In reality, individuals belong to different social niches and are therefore exposed to a potentially wide variety of audiences. For example, individuals can simultaneously belong to a professional occupation, a book club, a sports club, a musical band and/or an online gaming community. While formalizing this more realistic state of affairs (with n receivers instead of two) adds complexity, I expect the model's main predictions to remain the same.

[9] We assume that d_{sj} never takes a value of 0.

[10] For evidence that personality traits influence occupational choice, see Cobb-Clark & Tan (2011), De Fruyt & Mervielde (1999), Wells et al. (2016); for evidence that personality traits influence the activities that individuals indulge in, see Carlo et al. (2005) or Ickes et al. (1997). For evidence that a fit between individual traits and the niche's ideal trait profile is beneficial, see Denissen et al. (2018).

receiver r_j Player 3. A strategy for the sender in this game must specify (i) which social niche j she decides to join, (ii) which signal to send (or, alternatively, which social identity to adopt) in the signaling stage (either I_{g_m} or I_{g_j}) and (iii) whether to play $ALLC$ or $ALLD$ in her game with r_m and r_j (e.g., $ALLC_mALLD_j$, implying that the sender plays $ALLC$ with r_m but $ALLD$ with r_j, and written C_mD_j for convenience). An example of a strategy profile for the sender would be $\{j^*, I_{g_m}, C_mD_j\}$, where the sender would choose the social niche j^*, adopt the identity I_{g_m}, and cooperate with r_m but defect with r_j. A strategy for r_i in this game must specify whether to play $ALLC$ or $ALLD$ as a function of the sender's decision in the signaling stage. An example of a strategy for the receiver would be $ALLD_mALLD_j$ (written D_mD_j for convenience), where the sender would defect no matter what signal has been sent by the sender.

Proposition 3.4.2. *The strategy profile* $\{\{j^*, I_{g_m}, C_mD_j\}, C_mD_j, D_mC_j\}$ *is a SPNE of the game if the following conditions are satisfied:*

1. $j^* \in \underset{j \in J}{argmax}\, \bar{v}(j)$,
2. $k_m \geq k_j$,
3. $k_m \geq c$,
4. $\delta_m^h \geq \frac{c}{b}$.

At this equilibrium strategy profile, the sender adopts the identity I_{g_m} of receiver r_m and plays $ALLC$ (cooperates) only with r_m. Receiver m cooperates with the sender s if and only if the sender adopts I_{g_m}, while receiver j refuses to invest in a relationship with the sender if the sender adopts I_{g_m}.

Proposition 3.4.3. *The strategy profile* $\{\{j^*, I_{g_j}, D_mC_j\}, C_mD_j, D_mC_j\}$ *is a SPNE of this game if the following conditions are satisfied:*

1. $j^* \in \underset{j \in J}{argmax}\, \bar{v}(j)$,
2. $k_j \geq k_m$,
3. $k_j \geq c$,
4. $\delta_j^h \geq \frac{c}{b}$.

At this equilibrium strategy profile, the sender adopts the identity I_{g_j} of receiver r_j and plays $ALLC$ (cooperates) only with r_j. Receiver j cooperates with the sender s if and only if the sender adopts I_{g_j}, while receiver m refuses to invest in a relationship with the sender if the sender adopts I_{g_j}.

3.5 Discussion

This section revisits the puzzles of social identity through the lens of the predictions made by the theory developed in this chapter. I argue that (i) viewing social identity as a signal of trustworthiness and (ii) viewing individuals as adopting the social identity that can provide them with the greatest amount of (social) benefits, helps explain the puzzles of social identity.[11]

3.5.1 Puzzle 1: Social identity can be highly malleable

One prediction of the model developed in Section 3.4 is that since social identity signals trustworthiness in polarized environments, it will principally respond to social incentives. The argument exposed in this chapter is that social incentives principally take the form of long-term mutually beneficial relationships with other members of an individual's community. Therefore, if (i) an individual's community remains stable across time and (ii) the beliefs, ideologies and values adopted by community members remain stable, then the individual's social identity is expected to remain stable too. On the contrary, if (i) an individual's community changes, and/or (ii) the beliefs, ideologies and values adopted by community members change, then the individual's social identity is expected to change too.

I argue that the malleability of social identity stems from its signaling function in polarized environments. If the beliefs and values of other members of our social group or social networks change and we still want to appear on their side (trustworthy member of the community), then our social identity is expected to change too.

3.5.2 Puzzle 2: Social identity is environment-dependent

A second prediction of the theory is that individuals are expected to trade-off the benefits from cooperating with different audiences (or communities) when deciding which social identity to adopt. Formally, if k_m, the benefits that home community

[11] See Appendix 1 for a case study of religious beliefs among academic scientists which further illustrates the ideas developed in this chapter.

members can provide to the sender, are large relative to k_j, the benefits that the sender might receive from cooperating with members outside of one's home community, then it is expected that the sender will hold on her home community social identity and conversely if k_j is large relative to k_m.

I argue that the differential benefits from cooperating with different communities underlie the environment-dependency of social identity. When individuals join college (Lazer et al. 2010, Mayrl & Uecker 2011), a new workplace (Mutz & Mondak 2006) or a new neighborhood (Martin & Webster 2020, Sinclair 2012), other students, co-workers or neighbors can become the people with which they now interact the most and on which they now rely the most. These changes in social environment can shift the ratio of benefits that different audiences can provide and therefore modify the existing social incentives that individuals face.

3.5.3 Puzzle 3: Social identity can be resistant to conflicting evidence

A third prediction from the theory is that individuals will often want to hold on, and defend their social identity. Given that social identity serves as a signal of trustworthiness (or intention to cooperate) in polarized environments, individuals will be eager to make their social commitments public in order not to alienate other members of their community, especially in contexts in which the relevant aspects of social identity become particularly salient. In fact, if some beliefs and values have been adopted solely for their signaling value, then one does not expect individuals to modify their social identity when new (potentially conflicting) evidence arrives. Modifying their social identity in an environment in which it is common knowledge that social identity signals underlying social commitments ultimately amounts, in the eyes of receivers, to modifying their social commitments. This argument can explain why social identity accommodates positive information but appears resistant to conflicting evidence. Tellingly, Frimer et al. (2017) find that subjects consciously decide to avoid hearing non-congruent opinions by fear of undermining valuable relationships, which is exactly what the proposed theory predicts.

3.5.4 Puzzle 4: Social identity is correlated with genes and personality traits

A fourth prediction of the model developed in Section 3.4 is that individual-level traits (personality and cognition) can become correlated with specific beliefs and values. To see this, assume that there is only one relevant individual trait in the

sender's trait profile θ_s, p. That is, let $\theta_s = (p)$, with p taking one of two values: either $p = \underline{p}$, meaning that the sender has a low value for trait p, or $p = \bar{p}$, indicating that the sender has a high value for trait p. Moreover, assume that there are only two social niches to join, q and t. Social niche q's ideal trait profile is $\gamma_q = (\gamma_{q_p})$ and social niche t's ideal trait profile is $\gamma_t = (\gamma_{t_p})$. Assume that $\gamma_{q_p} = \underline{p}$ and $\gamma_{t_p} = \bar{p}$. Let the social group g_q be primarily represented among the members of q, while let the social group g_t be primarily represented among the members of t. The sender s with trait profile θ_s will choose the niche j providing her with the highest prospective value $\bar{v}(j)$, which is, by definition, the niche j^* that satisfies $j^* \in \underset{j \in J}{argmax}\, \bar{v}(j)$. Therefore, if s is endowed with $\theta_s = \underline{p}$, then s will choose to join q. Alternatively, if s is endowed with $\theta_s = \bar{p}$, then s will choose to join t. Given that members of q have primarily adopted the identity I_{g_q}, while members of t have primarily adopted I_{g_t}, a correlation endogenously arises between individual traits (here, p) and social identities, defined as packages of beliefs and values (here, I_{g_q} and I_{g_t}). At the aggregate level, we can then expect to observe a correlation between specific individual-level traits (e.g., personality traits) and specific beliefs and values. Since personality traits are highly heritable (Mitchell 2020), a correlation between genes and beliefs and values is also predicted.

While researchers usually assume that individuals endowed with different traits (or different psychologies or genes) respond differently to different beliefs or ideologies (Funk et al. 2013), the model outlined in this chapter predicts that the relationship between individual-level traits (personality, psychology or genes) and social identities is purely correlational. As such, the model helps explain some empirical findings that are not easily reconcilable with the idea that dispositional traits causally influence the adoption of specific beliefs and values. For instance, it helps to explain the finding that "genetic influence [on political attitudes] is manifest only after moving away from the parental home" (Hatemi et al. 2009, p.1153), meaning that the observed statistical relationship between genes/personality and beliefs and values only emerges once individuals leave their home community (Hatemi et al. 2009, Hufer et al. 2020). This is predicted by the model developed in this chapter. During development, psychologically and genetically dissimilar individuals are embedded into families and communities that strongly influence their beliefs and values. This will translate into the *shared environment* explaining most of the variance in beliefs and values before entering into adulthood. But once individuals have left their home, they usually need not hold on the beliefs of their parents, family or previous community anymore. Depending on the environment (social niche) they decide to join, they will congregate with genetically and psychologically similar oth-

ers and come to adopt new beliefs and values that are associated with the social groups other social niche members belong to, therefore explaining the finding that the correlation between genes and political attitudes only emerges once individuals have left their home.

3.6 Conclusion

This chapter has been motivated by the observation that social identity presents puzzles. It can be highly malleable but also resistant to information. It is environment-dependent but also correlated with genes and individual-level traits such as personality. In order to tackle these puzzles, I proposed to (i) view social identity as a signal of trustworthiness (or intention to cooperate) and (ii) view individuals as adopting the social identity that provides them with the greatest amount of (social) benefits.

The theory developed in this chapter makes several predictions, which can help explain the puzzles of social identity. First, if social identity signals trustworthiness in polarized environments, then individuals will be eager to adopt the same beliefs and values as others in their social group or social networks, so as to appear as trustworthy. This can explain the malleability of social identity. Second, individuals will have to trade off the benefits from cooperating with different individuals when deciding which social identity to adopt. This can explain changes in social identity when individuals join a new social environment, such as a new college, a new workplace or a new neighborhood. Third, if social identity is a signal of trustworthiness, then individuals might want to hold on and defend their social identity, so as to make their social commitments public. This can explain the resistance of social identity to conflicting evidence. Finally, in our theoretical model, individuals with similar traits join the same social niches. Research has shown that different beliefs and values (social identities) are more or less prevalent in different niches (Bonica 2014). I argue that the assortment of similar individuals in similar niches in which different beliefs and values are more or less enforced can explain the existing correlation between social identity and genes and personality.

3.7 Appendix 1: Case study on religious beliefs among academic scientists

A vibrant scientific movement has sought to understand the natural and cultural origins of religious thought (Barrett 2000, Norenzayan 2013). Some researchers have

argued that humans are cognitively pre-equipped for religious thought (Boyer 2001), while others have argued that the cultural transmission of religious thought and behavior was essential in sustaining large-scale cooperation (Norenzayan et al. 2016). Regardless of the origins of religious thought, religious and scientific beliefs are often pitted against each other, with religious beliefs being described as the quintessential opposite of scientific thought, sometimes even called delusions (Dawkins 2006). Even though the latter view is an extreme one, not shared by the vast majority of scientists, there is this sense, at least in Western society, that religious and scientific beliefs can not be mutually compatible—how can one believe without any evidence? I argue here that there appears to be a paradox only if one does not consider one of the main function of religious belief, which is of signaling trustworthiness.

Ecklund and colleagues have done extensive work on religious belief among academic scientists in the U.S. (Ecklund & Park 2009, Ecklund & Scheitle 2007) and cross-country (Bolger et al. 2019, Ecklund et al. 2016). I believe that their findings give credence to the view that religious belief can serve signaling purposes. First, their findings argue against the secularizing force of scientific training; that is, academics do not automatically abandon their religious beliefs once they have been scientifically trained. They find, for instance, that in India, 94% of scientists identify with a religious tradition, compared to only 30% in France (Ecklund et al. 2016, p.3). Moreover, in India, Hong-Kong and Taiwan, compared to the local population, a higher proportion of scientists "participate regularly in religious services" (Ecklund et al. 2016, p.4). This observation suggests that religious beliefs do not respond to scientific training; in other words, they do not respond to evidence (or lack thereof). It follows that when we observe relatively low levels of religious affiliation among academics in Western countries, we must wonder what, if not scientific training, is driving this pattern. I believe that the answer lies in the specific form academic norms have taken in Western countries.[12] Academics in Western countries are more likely to observe a conflict between religious and scientific beliefs, compared to countries like India or Taiwan (Ecklund et al. 2016, p.6).[13] As such, if they expect their colleagues to shun them because they hold beliefs incompatible with their scientific activity, academic newcomers in Western countries face more pressure to abandon their religious beliefs. In fact, Ecklund & Park (2009) conclude their study by noting that peer attitudes towards religion are a significant predictor of whether or not scientists see religion and science in conflict. This is in line with

[12] Ecklund & Scheitle (2007) also note that there is an important role for selection, in the sense that, at least in the U.S., non-religious disproportionately self-select into academia.

[13] Ecklund & Park (2009) find that, in the U.S., only 23% of academics believe that their colleagues have a positive view of religion.

Mayrl & Uecker (2011, p.181)'s finding that "[c]hange in religious beliefs appears [...] to be more strongly associated with network effects" among college students. This suggests that the maintenance or abandonment of religious beliefs is a direct function of the social environment in which individuals are embedded. Hence, religious beliefs seem to respond to social incentives, the incentive being smooth integration and cooperation with other members of the community one has decided to join.

Second, the findings of this group of researchers concerning academics who hold on to their religious beliefs are also instructive. By far the strongest predictor of religious affiliation among academics is religiosity in the home as a child. Ecklund & Scheitle (2007, p.301) find that, in the U.S., academics for whom religion was important in their family when growing up are "less likely to say that they currently do not see truth in religion", while Bolger et al. (2019, p.16) write that "religiosity at 16 was the single strongest predictor of current religiosity" in India, Italy and the United States. Can it be the case that, because they have been raised in religious families, these academics know more about religion, therefore explaining why they hold on their beliefs? We know from above that this can not be the case, because religious belief does not respond to facts or (lack of) evidence. I want to suggest here, in line with the theory developed in this chapter, that because they have been raised in an environment in which religion was important, these academics might still rely on a network (family, friends, previous community members) that is predominantly religious, therefore weakening the incentives of abandoning religious beliefs (k_m is high). This argument is consistent with the finding that academics raised in homes in which religion was not important are more likely to observe a conflict between religious and scientific beliefs (Ecklund & Park 2009). It is also consistent with findings that stress the primordial role of exposure to religious displays in religious belief acquisition (Gervais et al. 2021, Lanman 2012, Willard & Cingl 2017). Gervais et al. (2021, p.1374) found that "witnessing fewer credible displays of faith proved to be by far the most powerful predictor of religious disbelief". This suggests that when no one in your family, or among your friends and community members, displays a religious affiliation, you are unlikely to display one yourself. This is fully consistent with the idea that one will develop (or hold on) religious beliefs not because of their veracity or truth-value but because others in one's direct social environment have developed or have held on them. This argument, again, suggests that religious beliefs mainly respond to social incentives and can be used as a badge of trustworthiness (McCullough et al. 2016).

3.8 Appendix 2

Proof of Proposition 3.4.1. To see why combinations of C and D can not be part of a SPNE of the game, first note that this possibility could only happen when $\delta_i \geq \frac{c}{b}$ and $k_i \geq c$ both realize, given that we have determined that the only SPNE of the game when $\delta_i \geq \frac{c}{b}$ and/or $k_i \geq c$ do not realize is for both players to play $ALLD$. Assume that s plays D at one of her decision node x. The logic of SPNE therefore requires that r_i also plays D at his $(x-2)$ decision node, otherwise he will pay the costs of cooperation c without any further benefits. Similarly, if r_i plays D at one of his decision node x (which is not the initial node), then the logic of SPNE requires s to play D at her $(x-1)$ decision node. Therefore, if a player plays D at a decision node x, then both players necessarily play D until this decision node is attained at a SPNE of the game. Now, assume that s plays C at one of her decision node y. Given that $\delta_i \geq \frac{c}{b}$ holds, then r_i will also play C at his $(y-2)$ decision node. Similarly, if r_i plays C at one of his decision node y (which is not the initial node), then s will also play C at her $(y-1)$ node, given that $k_i \geq c$ holds. Therefore, if a player plays C at a decision node y, then both players necessarily play C until this decision node is attained at a SPNE of the game, when $\delta_i \geq \frac{c}{b}$ and $k_i \geq c$ both realize. It follows that at a SPNE of the game, either players play ALLD, or they play ALLC (when $\delta_i \geq \frac{c}{b}$ and $k_i \geq c$ both hold). ■

Chapter 4

The Persuasive Function of Positive Illusions

Summary

The objective of this chapter is to investigate whether positive illusions can persuade at equilibrium. In a two-player "Partner Choice" game, I show that "Low" types can pool with "High" types by adopting positive illusions. The equilibrium size of the illusion is predicted to be sensitive to the reputational costs of lying and the degree of observability of the Sender's underlying quality. In a three-player "Community" game, I show that positive illusions can remain stable provided that the equilibrium size of the lie is small enough, providing enough plausible deniability to the Sender. In both models, positive illusions are stable at equilibrium even though Receivers correctly anticipate the average value of the Sender's type. The empirical literature on positive illusions appears to support the main predictions of both models.

Classification

JEL Classification: C72, D82, D83

Keywords: Lying, Observability, Persuasion, Positive Illusions, Signaling

4.1 Introduction

A large literature has revealed that individuals often hold enhanced views (hereafter *positive illusions*) about themselves in a wide variety of domains. As described by McKay & Dennett (2009, p.505), "[s]uch illusions include unrealistically positive self-evaluations, exaggerated perceptions of personal control or mastery, and unrealistic optimism about the future". These tendencies often reflect themselves in individuals believing that they are better than average—particularly so on desirable traits—a phenomenon called the *better-than-average effect* (BTAE) (Guenther & Alicke 2010). These tendencies also reflect themselves in individuals often being *overconfident* about their abilities (Moore & Healy 2008). The commonality of such inaccurate perceptions has been a puzzle for researchers given the potentially high costs individuals might incur as a result of having inaccurate perceptions about themselves (Barber & Odean 2001, Baumeister et al. 1993, Bénabou & Tirole 2002, Fenton-O'Creevy et al. 2003).

Three main arguments have been developed to explain the stability of positive illusions. The first argument is that positive illusions provide individuals with *psychological and health benefits* (Taylor & Brown 1988). The idea is that holding accurate representations about oneself can be detrimental to psychological well-being, and that positive illusions can therefore help individuals cope with reality. The second argument is that positive illusions result from *cognitive or informational biases* (Moore & Healy 2008). Here, the main idea is that overly positive views about oneself result from individuals processing information in a non-motivated biased manner. In this view, positive illusions are therefore just errors individuals make in forming beliefs about themselves. The third argument, which is the subject of this chapter, is that individuals adopt positive illusions mainly to persuade others about their abilities (Kurzban & Aktipis 2007). The principal idea is that individuals might be better able to convince others of their (enhanced) abilities by holding enhanced beliefs about themselves.

Now, while this last argument is gaining traction, a severe flaw remains. If individuals adopt positive illusions to persuade others about their (enhanced) abilities, then why would others pay attention to this signal? In other words, if individuals expect others to enhance their representations of themselves, how can the signal (i.e., the positive illusion) still be informative? For a signaling system to remain stable, it must be honest on average (Johnstone & Grafen 1993); receivers can not be systematically fooled by senders, otherwise they will refrain from considering the signal (Frey & Voland 2011, Marshall et al. 2013). This necessary implies that for

the signaling system to remain stable, it must be beneficial (on average) for the receivers to adjust their behavior as a function of the signal. This point indicates that the signal must be at least partly informative. But how can the signal be informative if all senders send an enhanced signal about their abilities?

The main objective of this chapter is to investigate whether positive illusions can effectively persuade at equilibrium in a strategic communication game, embedded in a "Partner Choice" setting. That is, in a setting in which a Receiver (R) accepts a Sender (S) only if S is of higher quality, can S persuade R to accept her by adopting positive illusions (i.e., enhanced beliefs) about her type even though her type is lower than R's? I first show that when the payoffs of the interaction are deterministic, positive illusions can not persuade at equilibrium for the simple reason that when payoffs are deterministic, lies are immediately spotted by R. In such a context, if the reputational costs of lying are high enough, then all S types will send an honest signal about their quality at equilibrium; if the reputational costs of lying are too low, then S's signal becomes uninformative at equilibrium. If, however, the payoffs of the interaction are non-deterministic, then R can not infer with certainty S's true type from the payoffs. I show that this uncertainty allows "Low" types to pool with "High" types by presenting themselves as better than they are (by adopting positive illusions). Yet, R is not fooled and correctly infers the expected value of S's type. This implies that positive illusions can persuade, but not fool R at equilibrium. I also show that the range of "Low" types which can pool with "High" types at equilibrium is constrained by the reputational costs of lying and the degree of observability of S's type. Therefore, the "maximum size" of the illusion at equilibrium will decrease, the greater the reputational costs of lying and the greater the degree of observability of S's type.

In a second model, I investigate the stability of positive illusions in a group setting, with one Sender (S) and two Receivers (R_1 and R_2). The objective is to go beyond a two-player setting in order to analyze the conditions under which positive illusions can remain stable when an audience whose objective is to coordinate its response receives a signal about S's type. In this setting, S's objective is to be seen as better than she is and both Rs have to decide whether to punish S (or not) after observing the result of a task undertaken by S which depends on her true type. Both Rs want to punish S if and only if they expect the other to punish too. I show that in such settings, uncertainty about the other R's "punishment threshold" can allow positive illusions to remain stable at equilibrium provided that the size of the lie is sufficiently small. Therefore, "small lies" can remain unpunished since they can prevent coordinated punishment by the audience. This result suggests that when S

can plausibly deny having sending a lie, punishment can fail to be triggered, thereby stabilizing positive illusions at equilibrium.

The results of the theoretical analysis suggest that positive illusions will be sensitive to the degree of observability of the underlying quality, the reputational costs of lying and the ease with which individuals can plausibly deny having lied. In the last section of the chapter, I review the large literature on positive illusions through the lens of our theoretical analysis and try to contrast our predictions with those of alternative accounts. Specifically, I will contrast the persuasive account of positive illusions with the two main alternative models that have been proposed, namely positive illusions as helping individuals to cope with reality and positive illusions as cognitive biases. The first model predicts that positive illusions should not be sensitive to the degree of observability of the underlying quality, while the latter model predicts that positive illusions should be sensitive to neither the degree of observability nor the desirability of the underlying quality. By contrast, the theoretical analysis undertaken in this chapter predicts that positive illusions will be sensitive to both the desirability and the degree of observability of the underlying quality. I argue that the empirical evidence is consistent with the predictions of the persuasive account of positive illusions in the sense that positive illusions seem to be sensitive to the (social) desirability of the trait, the ease with which others can verify whether we are endowed with the advertised level of the trait and the ease with which individuals can plausibly deny having self-enhanced. Therefore, the persuasive account of positive illusions appears to be theoretically sound and empirically supported.

4.2 The disputed origins of positive illusions

This section will be dedicated to rapidly reviewing the three main arguments that have been developed to explain the stability of positive illusions and to outline their respective predictions.

4.2.1 Positive illusions and well-being

In their seminal paper, Taylor & Brown (1988) argue that positive illusions promote mental health. They reject the view that illusions reflect failures in information processing and rather defend the idea that positive illusions are individually adaptive under a variety of circumstances. More specifically, they argue the following:

> The individual who responds to negative, ambiguous, or unsupportive

> feedback with a positive sense of self, a belief in personal efficacy, and an optimistic sense of the future will, we maintain, be happier, more caring, and more productive than the individual who perceives this same information accurately and integrates it into his or her view of the self, the world, and the future. (Taylor & Brown 1988, p.205)

The argument is that having accurate representations about themselves will make individuals worse-off (in terms of well-being) and that they should, as a consequence, develop and maintain positive illusions in order to cope with reality. This argument is also supported by Baumeister (1989, p.188), who writes that "[i]t is depressing and maladaptive to see oneself too accurately" and by Greenberg et al. (1986, p.206), who write that "self-esteem gives people a basic sense of security that is needed very badly". This argument has also recently been defended by behavioral economists. For instance, Bénabou & Tirole (2016, p.146) write that "seeing oneself as smart, attractive, and good is intrinsically more satisfying than the reverse" while Loewenstein & Molnar (2018, p.1) note that some "beliefs are, in and of themselves, pleasurable to their holder". According to this line of thought, individuals adopt positive illusions as a way to improve their mental health, promote their well-being, protect their self-esteem and cope with the hardships of everyday life.

Evidence suggests that there exists a positive relationship between positive illusions and physical and mental health (Segerstrom et al. 1998, Taylor et al. 2000, 2003). For instance, Taylor et al. (2003, p.605) found that "high self-enhancers had lower cardiovascular responses to stress, more rapid cardiovascular recovery, and lower baseline cortisol levels". Furthermore, McKay & Dennett (2009) review a range of studies demonstrating the positive health effects of maintaining positive illusions, arguing that positive illusions are *evolved misbeliefs*. Now, while the link between positive illusions and mental and physical health exists, it is not clear whether this relationship is causal (Aspinwall & Tedeschi 2010).

While we might expect positive illusions to be upwardly unbounded to maximize psychological health, Baumeister (1989) has argued that there exists an *optimal margin of illusion*, beyond which illusions start to be harmful to the individual (in terms of costs due to bad decisions taken, based on the illusion). Therefore, we should expect individuals to hold enhanced beliefs about themselves but not unboundedly so, which is what we tend to observe (Sedikides & Gregg 2008). Now, the perspective that positive illusions are adopted solely to improve psychological health is, at its heart, individualistic. That is, standard explanations for why individuals adopt positive illusions focus primarily on the individual (Williams 2020). This implies that, from this perspective, we do not expect positive illusions to be

affected by variables such as the *observability* of the trait being enhanced, which describes the ease with which others can verify whether we are effectively endowed with the advertised level of the trait. This prediction will be confronted with empirical data in Section 4.8.

4.2.2 Positive illusions and cognitive and information-processing biases

A second argument that has been advanced in the literature on positive illusions is that the latter stem principally from cognitive and information-processing biases. That is, the argument is that positive illusions are not sustained by individual motivations to view oneself positively but rather result from individuals processing information in a biased manner. Summarizing the argument, Brown (2012) writes:

> [I]t has been suggested that informational differences (i.e., a tendency to know more about oneself than others), focalism (i.e., a tendency to focus on oneself when making comparative judgments), naive realism (i.e., a tendency to assume one's view of the world is a passive reflection of the world as it actually is), and egocentrism (i.e., a tendency to give undue weight to one's own perspective) produce a [BTAE] effect in the absence of any motivated need. (Brown 2012, p.210)

For instance, according to Kruger (1999), the BTAE effect stems from the fact that when people compare their abilities with those of others, their own level of ability serves as a judgmental anchor which is not sufficiently adjusted when evaluating the skills of others, ultimately generating an upward bias in comparisons. Chambers & Windschitl (2004, p.829), in a review of the literature, defend the idea that "there are numerous ways in which mechanisms that are not biased by self-enhancing motivations can nonetheless yield above-average and comparative-optimism effects", among which egocentrism, focalism and anchoring figure prominently. Moreover, Moore & Healy (2008) describe how phenomena such as overconfidence might arise due to the lack of information individuals have about their own performances, but especially about the performance of others. This lack of information, coupled with biases in inference, can lead subjects to overestimate their performance relative to the performance of others.

From this perspective, individual motivations are irrelevant in explaining positive illusions. The latter solely arise from individual cognitive and information-processing biases. It follows that the main prediction is that the trait or domain under consideration should be irrelevant to the nature and form of positive illusions.

That is, if individual motivations do not play a role in sustaining positive illusions, then the particular form that positive illusions take (direction, size, etc.) should be independent from the desirability of the trait under consideration (Brown 2012). Additionally, just as for the psychological account described above, the observability of the trait is predicted to have no influence on the nature of the positive illusion. These predictions will also be confronted with empirical data in Section 4.8.

4.2.3 Positive illusions and persuasion

The last argument reviewed here, which is the main focus of this chapter, is that individuals principally adopt positive illusions in order to influence the beliefs (and behavior) of others. More specifically, the argument is that by holding enhanced beliefs about themselves, individuals might be able to convince others of their (enhanced) abilities, ultimately influencing how others behave towards them. This argument is summarized by Kurzban & Aktipis (2007):

> If others can be made to believe that one is healthy, in control, and has a bright future, then one gains in value as a potential mate, exchange partner, and ally because of one's ability to generate positive reciprocal benefits in the future ... To the extent that there has been a history of competition for filling these social roles ... selection would have favored mechanisms that caused one to be convincing—without straining others' credulity—about being a good candidate to fill them. (Kurzban & Aktipis 2007, p.137)

Therefore, according to this argument, the stability of positive illusions stems from their interpersonal effects. If, by adopting positive illusions, individuals are better able to convince others to behave favorably towards them (e.g., accept as partners/allies/friends, attribute status, show deference, etc.), then individuals are better off compared to a situation in which they display accurate assessments of their abilities. This argument has been most famously defended by Von Hippel & Trivers (2011) in their theory of self-deception, writing that believing that one is better than one really is "can help us convince others that we are better (e.g., more moral, stronger, smarter) than we really are" (Von Hippel & Trivers 2011, p.4).

Empirical evidence for the persuasive effects of positive illusions has recently begun to emerge. For instance, Anderson et al. (2012, p.730) have shown, in their study, that "overconfident individuals were perceived by others as more competent and, in turn, afforded higher status", while Lamba & Nityananda (2014, p.4) describe how "[o]verconfident individuals were overrated and underconfident individuals were

underrated", suggesting that displaying overconfidence improves our image in the eyes of others. Recent experimental studies have also shown that subjects tend to (unconsciously) become overconfident when placed in contexts in which the objective is to persuade others about some trait, and this tends to be individually beneficial, given that other subjects attend to this level of confidence when deciding how to behave (Charness et al. 2018, Schwardmann & Van der Weele 2019, Schwardmann et al. 2022, Solda et al. 2019). These studies therefore suggest that in social interactions, it can be beneficial for individuals to enhance their views about themselves due to the persuasive effects that this might have on others.

As discussed in the Introduction, while this argument is gaining traction, it is unclear how positive illusions can persuade at equilibrium. If the persuasive account of positive illusions is right, then what keeps positive illusions informative? What are the mechanisms that might contribute to the stability of positive illusions? In Section 4.5 and Section 4.6, I will present two models in which I try to answer these questions. The main objective of the first, "Partner Choice" model, will be to investigate whether positive illusions can effectively persuade at equilibrium; that is, whether a Sender (S) can persuade a Receiver (R) to accept her as a partner even though she is of lower quality. This would give credence to the argument exposed above. If such an equilibrium exists, a second objective will be to precisely outline its nature and characteristics in order to make predictions about the conditions under which it might be observed. The main objective of the second, three-player "Community" game, with one Sender (S) and two Receivers (R_1 and R_2), will be to investigate the stability of positive illusions in group settings. More specifically, we will analyze the role of plausible deniability (i.e., "small" lies) in preventing coordinated punishment. Before delving into the detailed description of the models, we will first survey the literature on which we build the present theoretical analysis and then discuss how the models need to be interpreted.

4.3 Relevant theoretical literature

We build on the large literature on games of *strategic communication*, which are games in which a Sender (S) has private information about the state of the world and can send messages in order to convey information to an uninformed Receiver (R). In their seminal "cheap-talk" paper, Crawford & Sobel (1982) have shown that the smaller the conflict of interest between S and R, the more information can be transmitted at equilibrium. At the extreme, when the conflict of interest is largest, messages become completely uninformative. Their paper has therefore highlighted

the need for some mechanism—here, shared interests or shared preferences—to keep messages informative at equilibrium. Another mechanism, which is the topic of this chapter, is the (reputational) cost of lying about the true state of the world. Reputational costs of lying are ubiquitous in humans and other animals and they have been shown to help stabilize signaling systems, keeping messages informative at equilibrium (Lachmann et al. 2001, Webster et al. 2018). Evolutionary models have in fact confirmed that repeated interactions can, via reputation, maintain honesty at equilibrium (Rich & Zollman 2016, Silk et al. 2000). Yet, while these models are informative about the conditions under which honesty can evolve, they tend (for tractability) to severely restrict the range of strategies that players can use.

Closest to our setting, costs of lying (defined as misrepresentation of private information) have been incorporated in a strategic communication setting by Kartik (2009), whose paper blends the literature on cheap talk, costly signaling and verifiable disclosure. In his paper, an upwardly biased Sender (S) has private information about her type and sends a message about her type to an uninformed Receiver (R) who has to take an action. S's objective is to persuade R that her type is higher than it actually is. Kartik (2009) assumes that S incurs an exogenous cost of lying which increases with the size of the lie. He shows that inflated language naturally arises at equilibrium when lying costs are not too high, with (almost) all S types claiming to be of higher type than they actually are. While R recognizes that S is using inflated language at equilibrium, it is still in S's interest to send an enhanced signal about her quality, exactly because R expects this: sending an honest signal would make R infer that S's type is in fact *lower* than it actually is. In this important paper, Kartik (2009) therefore shows that inflated messages—which can be interpreted as positive illusions—are expected at equilibrium when the underlying quality is private information, when lying is moderately costly and Senders wish to persuade Receivers that they are of higher quality than they actually are. Yet, this paper does not answer the question which is the subject of the present chapter: can positive illusions persuade R to accept S at equilibrium? In order to answer this question, we will integrate the strategic communication game in a "Partner Choice" setting in Section 4.5. Additionally, while Kartik (2009) does not consider the observability of the Sender's type, this parameter will play a crucial role in our model.

Another point of departure from Kartik (2009)'s paper concerns the way we are going to formalize lying costs. While Kartik (2009) assumes that lies are costly, with the cost increasing with the size of the lie, we will follow Dziuda & Salas (2018) and Balbuzanov (2019)'s approach in assuming that lies are costless, but that greater

lies increase the probability of being caught lying and therefore the probability of incurring reputational costs. Yet, these papers assume an exogenous probability of lie detection, while we assume in this chapter that the probability of detection is a function of the size of the lie and the degree of observability of the Sender's type.

Finally, building on Khalmetski & Sliwka (2019)'s work, Fries et al. (2021) introduce (intrinsic and extrinsic) lying costs as well as a degree of observability of the lie in a cheating game. They show that the greater the degree of observability of the lie, the smaller the likelihood that agents will lie at equilibrium. While they formalize observability as the Sender's belief that the Receiver knows the true state of the world, we will in the present chapter distinguish between the *ex ante* and *ex post* degree of observability of the state of the world (i.e., the Sender's type). The *ex ante* degree of observability of the S's type will describe the ease with which R can infer S's type *prior* to the interaction. The *ex post* degree of observability of S's type will describe the ease with which R can infer S's type *after* the interaction upon observing the payoffs. We will show that this distinction has important ramifications for the stability of positive illusions in a "Partner Choice" setting and helps us better understand the empirical evidence surveyed in Section 4.8.

4.4 A note on the interpretation of the models

Although positive illusions are defined as enhanced *beliefs* about one's type, the following models do not formalize the process of belief formation about the sender's type. In fact, the models do not explicitly formalize the beliefs the sender has about her type. Rather, the models are static, the sender is assumed to know her underlying type and has to decide what signal (or message) to send about her type. The signal she sends is assumed to represent the beliefs she has about her type. A signal greater than her true type is interpreted as a *positive illusion* (an enhanced belief about her type). But if the sender is assumed to know her true type, then how can she simultaneously adopt positive illusions about her type? Moreover, can we reasonably say that she truly believes the signals that she sends about her type?

The way the models need to be interpreted is as follows. While the analysis is static, the equilibrium strategy profiles have to be considered as the end-products of a process of learning or evolution (Fudenberg & Levine 1998). The assumption that the sender knows her true type is not essential and meant only to facilitate the theoretical analysis. The sender might start the process having accurate beliefs about her type, but her beliefs will converge over time (through learning or evolution) towards the ones that maximize her payoffs in her strategic interactions with the

receiver(s). The equilibrium belief will usually be a function of the sender's true type; what ultimately is of interest is the connection between the sender's true type and her equilibrium belief. This process of learning or evolution of course happens subconsciously (the sender is not consciously deciding which beliefs she adopts about herself).[1] Therefore, the sender's equilibrium strategy, which describes the signal that she sends (hence the beliefs that she holds about herself) needs to be understood as the end result of a dynamic process converging towards optimality.

Yet, can we say that the sender really believes the signal that she sends? Can she not publicly pretend to be of higher type, while privately having accurate beliefs about her true type? Here, the argument is that the sender will internalize the beliefs that she is incentivized to adopt. Again, this process surely happens subconsciously and automatically. For instance, Melnikoff & Strohminger (2020) have shown that even lawyers, who are trained to prevent advocacy to bias their judgments, bias their beliefs when advocating for a cause. Therefore, if adopting enhanced beliefs (i.e., adopting *positive illusions*) about her type is payoff-maximizing and facilitates persuasion, then the sender can be expected to truly adopt such beliefs. This is the perspective I am taking in this chapter.

4.5 Two-player "Partner Choice" game

4.5.1 Model with deterministic payoffs

4.5.1.A Model setup

There are two players, a *Sender* (S) and a *Receiver* (R). The Receiver has underlying quality $\theta_R > 0$, which is known to R and perfectly observable by S.[2] The Sender has private information about her underlying quality θ_S, with θ_S uniformly distributed on $\Theta_S = [\theta_R - \gamma, \theta_R + \gamma]$, with $\gamma > 0$. The interpretation is the following: upon interacting with S, R is unsure about whether S is of higher or lower quality, and the higher the value of γ, the wider the range of qualities deemed plausible by R. While strictly speaking the cardinality of the set Θ_S is the same for all values of $\gamma > 0$, decreasing γ can be interpreted as decreasing R's uncertainty about the value of θ_S. Therefore, we interpret γ as representing the *ex ante degree of observability* of θ_S, with higher values of γ implying that θ_S is ex ante less observable (or verifiable) by

[1] For an illuminating discussion about how evolution and learning processes shape our beliefs and preferences, see Hoffman & Yoeli (2022, Chapter 2).

[2] Given that this chapter primarily focuses on S's presentation to R, this assumption is meant only to facilitate the ensuing analysis.

R upon interacting with S.[3]

In this chapter, we focus our analysis on S's *presentation* to R. More specifically, we assume that upon interacting with R, S sends a cost-free signal about θ_S, m_{θ_S}. The signal m_{θ_S} can take any value in $\Theta_S = [\theta_R - \gamma, \theta_R + \gamma]$. This signal is meant to represent the way S presents herself to R. If $m_{\theta_S} = \theta_S$, then S honestly represents her quality to R. However, if $m_{\theta_S} > \theta_S$, then S presents herself as better than she is. A signal $m_{\theta_S} > \theta_S$ will be called a *positive illusion*. This signal is assumed to be cost-free given that its production costs are negligible.

S's objective is to be accepted by R. If R accepts S, then both players engage in a joint project which generates payoffs π_i, with $i \in \{S, R\}$. If R denies S, then the game ends and both players have payoffs equal to ϕ_i, which is player i's outside option, with $i \in \{S, R\}$. On the one hand, we assume that S's payoff from interacting with R is strictly increasing with θ_R and (for simplicity) independent from θ_S, such that $\pi_S(\theta_S, \theta_R) = \phi_S + \beta\theta_R$, with $\beta > 0$. On the other hand, R wants to interact with S only if S has *higher* quality. That is, R wants to interact with S only if $\theta_S > \theta_R$. This assumption is translated in the following payoffs for R from interacting with S: $\pi_R(\theta_S, \theta_R) = \phi_R + \alpha(\theta_S - \theta_R)$, with $\alpha > 0$. The payoffs for R are such that: (i) R is indifferent between denying and accepting S if S has the same quality (if $\theta_S = \theta_R$),[4] (ii) R prefers to deny rather than to accept S if S has lower quality (if $\theta_S < \theta_R$) and (iii) R prefers to accept rather than to deny S if S has higher quality (if $\theta_S > \theta_R$).

When deciding whether to accept or deny S, R has to rely on his prior about θ_S as well as on S's signal m_{θ_S}. We assume that upon receiving m_{θ_S}, R forms an expectation $\mathbb{E}[\theta_S | m_{\theta_S}]$ about S's underlying quality, θ_S. This expectation translates into an expected payoff from interacting with S, namely: $\mathbb{E}_{\theta_S}[\pi_R] = \phi_R + \alpha(\mathbb{E}[\theta_S | m_{\theta_S}] - \theta_R)$.

We assume that S suffers reputational costs if S is caught lying about her underlying quality θ_S. Therefore, while the production costs of the signal m_{θ_S} are zero, a lie can be costly if spotted. As a consequence, while the game is formalized as being played only between S and R, we must interpret the present game as being embedded in a broader social setting (in which reputations are at stake), with strategies in this game potentially influencing individual payoffs in another (larger) game. Therefore, this interaction is not isolated. More precisely, we assume the following:

a. If R accepts S after receiving signal m_{θ_S} and the realized payoff is greater or equal than the expected payoff (i.e., if $\pi_R(\theta_S, \theta_R) \geq \mathbb{E}_{\theta_S}[\pi_R]$), then the game

[3] This interpretation will be useful when we contrast the main predictions of the model with empirical evidence in Section 4.8.

[4] We assume that if $\theta_S = \theta_R$, then R decides to deny S.

ends after payoffs are realized.[5]

b. If R accepts S after receiving signal m_{θ_S}, and the realized payoff is strictly lower than the expected payoff (i.e., if $\pi_R(\theta_S, \theta_R) < \mathbb{E}_{\theta_S}[\pi_R]$), then the game ends after payoffs are realized and S suffers a cost c due to R's inference that S has lied about θ_S.

To summarize, the timing of the game is as follows:

1. Upon meeting, S sends a cost-free signal $m_{\theta_S} \in \Theta_S$ about θ_S and R forms an expectation $\mathbb{E}[\theta_S|m_{\theta_S}]$ about S's underlying quality, θ_S, which translates into an expected payoff from interacting with S, $\mathbb{E}_{\theta_S}[\pi_R]$, as described above.

2. R decides whether to *Accept* (A) or *Deny* (D) the Sender. If R denies S, then the game ends and both players receive payoffs $\pi_i = \phi_i$, with $i \in \{S, R\}$. If R accepts S, then both players engage in a joint project which results in payoffs $\pi_i(\theta_S, \theta_R)$, with $i \in \{S, R\}$, as described above. If R's realized payoff is lower than the expected payoff (i.e., if $\pi_R(\theta_S, \theta_R) < \mathbb{E}_{\theta_S}[\pi_R]$), then S also suffers a cost c.

4.5.1.B Player strategies

In this game, S has to decide what cost-free signal m_{θ_S} to send given her underlying quality θ_S. Therefore, let $m : \Theta_S \to \Theta_S$ represent S's strategy. R has to decide whether to accept or deny S as a function of the signal m_{θ_S}. Therefore, let $\sigma : \Theta_S \to \{A, D\}$ represent R's strategy.

4.5.1.C Equilibrium concept

We solve for Perfect Bayesian Equilibrium (PBE) of the game, which requires that all S types best-respond to the strategy played by other types and to the strategy played by R, that R plays a best-response to S given his beliefs about θ_S and that R's beliefs are updated using Bayes' Rule on the equilibrium path.

Given the strategic setting (see Figure 4.1), we focus on semi-separating equilibria, which are equilibria with a cutoff value $\theta_S^* \in [\theta_R - \gamma, \theta_R + \gamma]$ such that all types $\theta_S < \theta_S^*$ send $m_{\theta_S}^l = \theta_S^*$ and are denied by R and all types $\theta_S \geq \theta_S^*$ send $m_{\theta_S}^h = \theta_R + \gamma$ and are accepted by R. We furthermore require that $\theta_S^* < \theta_R$ at equilibrium, given that if $\theta_S^* \geq \theta_R$, then positive illusions can not persuade. Therefore, we wonder whether "Low" types can pool with "High" types at equilibrium and by so doing be accepted by R.

[5] This assumption implies that R does not punish S if S sends $m_{\theta_S} < \theta_S$.

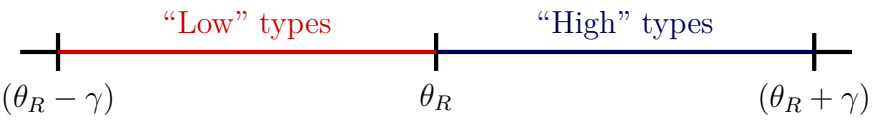


Figure 4.1 – Set Θ_S of S types. This set can be divided in two parts, with "Low" types being S types that R would want to deny and "High" types being S types that R would want to accept.

4.5.1.D Model resolution

In the present setting, can S persuade R to accept her by sending a signal $m_{\theta_S} > \theta_R$ (a positive illusion), even though $\theta_S < \theta_R$? Proposition 4.5.1 describes our first result. All proofs can be found in the Appendix.

Proposition 4.5.1. *When payoffs are deterministic, positive illusions can not persuade at equilibrium.*

The intuition behind Proposition 4.5.1 is very simple: if payoffs are deterministic, then any lie $m_{\theta_S} \neq \theta_S$ will be spotted by R with certainty. If the reputational costs of lying are high enough ($c \geq \beta\theta_R$), then S will refrain from sending a lie at equilibrium and there exists a *separating* equilibrium at which S sends $m^*_{\theta_S} = \theta_S$ for all $\theta_S \in \Theta_S$, and R accepts S if and only if $m^*_{\theta_S} > \theta_R$. If, on the contrary, the reputational costs of lying are too small ($c < \beta\theta_R$), then the signal m_{θ_S} becomes uninformative given that R can expect all S types to send the highest possible signal. In such a case, there exists a *babbling* equilibrium at which S sends $m^*_{\theta_S} = \theta_R + \gamma$, for all $\theta_S \in \Theta_S$, and R always denies S.

Therefore, when payoffs are deterministic, S can not take advantage of R's uncertainty about θ_S to present herself as better than she is and convince R to accept her. The prediction is that we should observe individuals having rather accurate beliefs about their underlying quality when (i) their underlying quality is easily observable ($\gamma \to 0$) and (ii) the outcome of tasks is easily predicted by the individual's underlying quality *and* the reputational costs of lying are high enough.

Now, we have been assuming that there exists a one-to-one correspondence between S's quality (θ_S) and R's payoff from interacting with S ($\pi_R(\theta_S, \theta_R)$). This assumption probably does not capture most situations of everyday life. More often than not, payoffs from joint projects, tasks or interactions are noisy, implying that R's payoff $\pi_R(\theta_S, \theta_R)$ might have been generated by a range of qualities θ_S. In such situations, observing $\pi_R(\theta_S, \theta_R)$ does not immediately allow R to infer θ_S with certainty. In the next section, we introduce noise in the payoffs from the joint project and investigate how this new element might affect the persuasiveness of positive illusions at equilibrium.

4.5.2 Model with non-deterministic payoffs

4.5.2.A Model setup

Let us assume that the timing of the game between S and R is the same as in Section 4.5.1.A. The only difference lies in the value of R's payoff from interacting with S ($\pi_R(\theta_S, \theta_R)$). More specifically, we assume that $\pi_S(\theta_S, \theta_R) = \phi_S + \beta\theta_R$, with $\beta > 0$ as before. However, we now set $\pi_R(\theta_S, \theta_R) = \phi_R + \alpha(\theta_S - \theta_R) + Z$, with $\alpha > 0$ and Z uniformly distributed on $[-z, z]$, independent of θ_S. It follows that upon interacting with S with underlying quality θ_S, payoffs in the interval $\Pi_R = [\phi_R + \alpha(\theta_S - \theta_R) - z, \phi_R + \alpha(\theta_S - \theta_R) + z]$ are *ex ante* equally likely for R. Therefore, higher values of z imply that it is more difficult for R to infer the underlying quality θ_S from the value of the payoffs. We will therefore interpret z as the *ex post degree of observability* of the underlying quality θ_S, with higher values of z implying that θ_S is ex post less observable (or verifiable) by R.[6]

As before, upon receiving m_{θ_S}, R forms an expectation $\mathbb{E}[\theta_S|m_{\theta_S}]$ about S's underlying quality, θ_S. This expectation translates into an expected payoff from interacting with S, $\mathbb{E}_{\theta_S}[\pi_R] = \phi_R + \alpha(\mathbb{E}[\theta_S|m_{\theta_S}] - \theta_R)$ given that $\mathbb{E}(Z) = 0$.

As before, we assume that S incurs a reputational cost c if she is caught lying about her underlying quality θ_S. More precisely, we assume the following:

a. If R accepts S after receiving signal m_{θ_S}, and the realized payoff is greater or equal than the lower bound of Π_R (greater or equal than $\phi_R + \alpha(m_{\theta_S} - \theta_R) - z$), then the game ends after payoffs are realized.

b. If R accepts S after receiving signal m_{θ_S}, and the realized payoff is strictly lower than the lower bound of Π_R (strictly lower than $\phi_R + \alpha(m_{\theta_S} - \theta_R) - z$), then the game ends after payoffs are realized and S suffers a cost c due to R's inference that S has lied about θ_S.

4.5.2.B Player strategies and equilibrium concept

As before, let $m : \Theta_S \to \Theta_S$ represent S's strategy and let $\sigma : \Theta_S \to \{A, D\}$ represent R's strategy. We solve for (semi-separating) Perfect Bayesian Equilibrium (PBE) of the game.

[6] While the parameter z is interpreted as the variability in the payoffs from the joint project, z can also be seen as the degree to which S can plausibly argue that the realized payoff, if lower than expected, actually stems from a higher (than warranted) quality. For instance, if S can plausibly argue that a low result on an IQ test derives from external factors (e.g., that the test was inadequate, that intelligence can not be measured via IQ tests, etc.), then S might avoid the reputational costs of being caught lying. That is, z can also be interpreted as the extent to which S can have recourse to auxiliary hypotheses (Gershman 2019) to justify lower than expected payoffs.

4.5.2.C Model resolution

Can noisy payoffs help positive illusions persuade? Proposition 4.5.2 describes the equilibrium cutoff value of θ_S^*, which represents the S type that is indifferent between sending θ_S^* (and being denied by R) and sending $\theta_R + \gamma$ (and being accepted by R).

Proposition 4.5.2. *For any semi-separating equilibrium, the cutoff value is given by* $\theta_S^* = (\theta_R + \gamma) - \frac{4z\beta\theta_R}{\alpha c}$.

One can see that the higher the value of z, the lower the value of θ_S^* at equilibrium, which suggests that reducing the (ex post) observability of θ_S can increase the range of S types which can send an enhanced signal about their quality at equilibrium. Alternatively, the higher the value of c, the higher the value of θ_S^* at equilibrium, suggesting that increasing the reputational costs of lying reduces the range of S types which can send an enhanced signal about their quality at equilibrium.

Note also that increasing θ_R increases the value of θ_S^* at equilibrium only if $c > \frac{4z\beta}{\alpha}$. This suggests that if the reputational costs of lying c are not high enough (particularly with respect to z) then increasing R's underlying quality θ_R will increase the range of S types which are willing to send an enhanced signal about their underlying quality at equilibrium. Finally, increasing the value of γ mechanically increases the value of θ_S^* at equilibrium.

Proposition 4.5.3 describes the range of values of c for which $\theta_S^* \in (\theta_R - \gamma, \theta_R)$ at equilibrium.

Proposition 4.5.3. *If* $\frac{2\theta_R z\beta}{\alpha\gamma} < c < \frac{4\theta_R z\beta}{\alpha\gamma}$, *then at equilibrium,* $\theta_S^* \in (\theta_R - \gamma, \theta_R)$.

On the one hand, if $c \leq \frac{2\theta_R z\beta}{\alpha\gamma}$, then the reputational costs of lying are too low to prevent all S types to send the highest possible signal $\theta_R + \gamma$. On the other hand, if $c \geq \frac{4\theta_R z\beta}{\alpha\gamma}$, then the reputational costs of lying are too high for S to be willing to send an enhanced signal about her quality at equilibrium, when $\theta_S < \theta_R$. Therefore, c must be neither too high nor too low for positive illusions to remain stable at equilibrium, and the higher the cost c, the smaller the range of θ_S willing to send an enhanced signal about their underlying quality.

We can also see that the smaller the (ex post) degree of observability of θ_S (the higher z is), the greater c needs to be at equilibrium. This result hints at the important role of reputational costs of lying when the underlying quality is not easily inferred from payoffs of joint projects, interactions, etc. Alternatively, this result suggests that reputational costs play a lesser role when z is low, that is, when the value of θ_S can easily be inferred (or verified) by R.

Let $m^l_{\theta_S} = \frac{\theta^*_S+(\theta_R-\gamma)}{2}$ = "*Low*" and $m^h_{\theta_S} = \frac{\theta^*_S+(\theta_R+\gamma)}{2}$ = "*High*". Proposition 4.5.4 describes our main equilibrium of interest.

Proposition 4.5.4. *There exists a semi-separating Perfect Bayesian Equilibrium of the game with cutoff value* $\theta^*_S = [(\theta_R+\gamma) - \frac{4z\beta\theta_R}{\alpha c}] \in (\theta_R-\gamma, \theta_R)$*, at which all types* $\theta_S < \theta^*_S$ *send* $m^{l^*}_{\theta_S}$ = "*Low*" *and are denied by R, and all types* $\theta_S \geq \theta^*_S$ *send* $m^{h^*}_{\theta_S}$ = "*High*" *and are accepted by R, as long as* $\frac{2\theta_R z\beta}{\alpha\gamma} < c < \frac{4\theta_R z\beta}{\alpha\gamma}$ *is satisfied.*

At the above-described equilibrium, some S types with $\theta_S < \theta_R$ can pool with high-quality types by sending the signal $m^{h^*}_{\theta_S}$ and, by doing so, can convince R to accept them (see Figure 4.2). One might conclude that positive illusions can in fact persuade at equilibrium. While it is true that R will accept some S with $\theta_S < \theta_R$, R correctly anticipates the average value of θ_S when receiving $m^{h^*}_{\theta_S}$ at the above-described equilibrium. Therefore R is not fooled by S and in fact necessarily benefits (on average) from accepting S when receiving $m^{h^*}_{\theta_S}$.

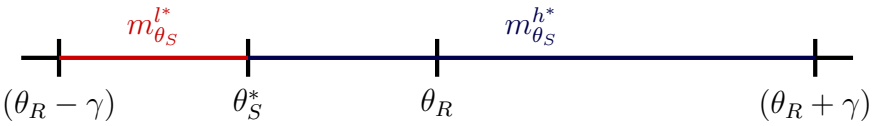


Figure 4.2 – Graphical representation of the equilibrium described in Proposition 4.5.4, with $\theta^*_S \in (\theta_R-\gamma, \theta_R)$.

4.5.2.D Discussion

The above analysis shows that when payoffs are non-deterministic, which implies here that R can not infer S's underlying quality θ_S with certainty upon observing the payoffs of the joint project, "Low" types can pool with "High" types by presenting themselves as better than they are (by adopting positive illusions). Therefore, some "Low" types can take advantage of the inherent uncertainty stemming from the lack of observability (high γ) and the difficulty in inferring (high z) S's underlying quality θ_S to pool with "High" types at equilibrium.

Adopting positive illusions can therefore help persuade, even though R can not be fooled at equilibrium. Given the threat of reputational costs of lying for S, R can be confident that he will benefit (on average) from accepting S when receiving the "High" type signal $m^{h^*}_{\theta_S}$. This is due to the fact that the reputational costs of lying c prevent extremely "Low" types to pool with "High" types at equilibrium by making it too risky for them (in expectation) to send the "High" type signal, and the higher the value of c, the smaller the range of "Low" types able to pool with "High" types at equilibrium.

If we interpret the equilibrium distance between θ_S^* and θ_R as the maximum size of the lie (the positive illusion), then increasing the value of c will result in decreasing the maximum size of the lie at equilibrium. On the other hand, increasing z will (ceteris paribus) have the opposite effect by allowing a greater range of "Low" types to pool with "High" types. Our analysis therefore underlines the important role of the (ex post) degree of observability of θ_S (z) as well as the reputational costs of lying (c) in the persuasiveness of positive illusions at equilibrium. We expect positive illusions to be greater, the smaller the reputational costs of lying c and the greater the noise in payoffs z.

Finally, a note about the interpretation of the two-player "Partner Choice" game: although the strategic interaction has been modeled as R deciding whether or not to accept S to undertake a joint project generating payoffs for both players as a function of their underlying quality, we can interpret the results in a more general way. Any situation in which observers (receivers) try to infer the underlying quality of a sender in order to decide how to behave towards them and receive noisy signals about this quality based on their impressions and results from a test, task, project, interaction, etc., can be interpreted through the lens of this model, and the main results are expected to extend to such situations. That is, the main predictions of the model are expected to apply to any context in which S benefits from being seen (overly) positively by R and R has to decide how to behave with respect to S (e.g., attribute status, accept in a group, compete with, gossip, etc.) based on noisy signals and feedback about the underlying quality θ_S.

4.6 Three-player "Community" game

The "Partner Choice" game in Section 4.5 has analyzed the stability of positive illusions in a two-player game. What about contexts in which a Sender interacts with an audience composed of several Receivers? Such settings are very common: at work, in a sports team, at a party, etc. Very often, audience members want to coordinate their responses, meaning they would want to punish S for self-enhancing if and only if they expect others to punish too (Boyd et al. 2010, Molleman et al. 2019, Panchanathan & Boyd 2004, Wiessner 2005). This necessarily implies that beliefs about what others believe (higher-order beliefs) are important in such settings (Hoffman et al. 2018).

The present setting is meant to highlight the role that Receivers' higher-order beliefs might play in contributing to the stability of positive illusions. In this new setting, assume that there is one sender (S) and an audience composed (for sim-

plicity) of two receivers (R_1 and R_2). Assume that both R_1 and R_2 share the same underlying quality, with $\theta_{R_1} = \theta_{R_2} = \theta_R$. As before, S has underlying quality θ_S which is uniformly distributed on $[\theta_R - \gamma, \theta_R + \gamma]$.

The timing of the game is as follows:

1. At the start of the interaction, S sends a *cost-free* signal about her underlying quality, $m_{\theta_S} \in \Theta_S = [\theta_R - \gamma, \theta_R + \gamma]$. After receiving m_{θ_S}, both receivers form an expectation $\mathbb{E}_i[\theta_S | m_{\theta_S}]$, with $i \in \{1, 2\}$, about S's underlying quality θ_S. Given that both receivers have the same information, we assume that $\mathbb{E}_1[\theta_S | m_{\theta_S}] = \mathbb{E}_2[\theta_S | m_{\theta_S}]$.

2. S engages in a task (which can be a test, a project, an interaction, etc.) generating results as a function of θ_S. We assume, for the purposes of this demonstration, that the task result is deterministic, such that the result of the task undertaken by S with quality θ_S is determined by $\pi_{\theta_S} = \theta_S$. While both R_1 and R_2 know the payoff function, we assume, without loss of generality, that R_1 is unsure about R_2's belief about π_{θ_S} (while R_2 knows that R_1 knows π_{θ_S}). That is, the only uncertainty lies in R_1's belief about R_2's belief about π_{θ_S}. More specifically, we assume that R_1 believes that, for R_2, upon interacting with S with quality θ_S, any task result in $\pi_{\theta_S} = [\theta_S - z, \theta_S + z]$ is *ex ante* equally likely, with z a random variable uniformly distributed on $[0, \delta]$, independent of θ_S. We assume that $\delta < \gamma$, meaning that R_1's uncertainty about R_2's beliefs about π_{θ_S} is bounded by the (ex ante) observability of θ_S. Therefore, R_1 believes that R_2 believes that a wider range of qualities might generate a given payoff π_{θ_S}. It is clear that for both receivers, the expected payoff of interacting with S with quality θ_S is equal to $\Pi_{R_i} = \pi_{\theta_S} = \theta_S$, $i \in \{1, 2\}$, although R_1 believes that for R_2, $\Pi_{R_2} = \pi_{\theta_S} = [\theta_S - z, \theta_S + z]$, with z a random variable uniformly distributed on $[0, \delta]$.

3. Upon observing the result of the task, both receivers play a coordination game. They can decide whether to *punish* S (play P) or *not punish* S (play $\bar{P}$). That is, both receivers would like to punish S *if and only if* they expect the other receiver to punish too. The payoff matrix of the coordination game is the following:

		R_2	
		P	$\bar{P}$
R_1	P	(x, x)	(m, n)
	$\bar{P}$	(n, m)	(y, y)

For it to be a coordination game, we need to set $x > n$ and $y > m$. The *risk-dominance* of the coordination game is written $\bar{q} = \frac{y-m}{(y-m)+(x-n)}$. The interpretation is that for R_1 to be willing to play P in the coordination game, he must expect R_2 to play P with a probability greater than $\bar{q}$, and conversely. It is assumed that if both receivers decide to punish S (play P) then S incurs a cost $c > 0$. If only one receiver decides to punish S then S incurs a cost v, with $0 < v < c$. If both receivers decide not to punish S (play $\bar{P}$) then S incurs no cost.

4. Payoffs are realized after both receivers play the coordination game. If both receivers play P, then S's payoff is equal to $\pi_S = m_{\theta_S} - c$. If both receivers play $\bar{P}$, then S's payoff is equal to $\pi_S = m_{\theta_S}$. Finally, if only one receiver plays P, then S's payoff is equal to $\pi_S = m_{\theta_S} - v$.

As before, we assume that both receivers are inclined to punish S if the result of the task is strictly lower than the lower bound of Π_{R_i}. From the setting described above, it is clear that both receivers would be inclined to punish S as long as $\pi_{\theta_S} = \theta_S < m_{\theta_S}$. This can be called R_i's "punishment threshold". What differentiates this setting from the one studied in Section 4.5 is that R_i will be willing to punish S only if he is sufficiently confident that the other receiver will punish too (he must be sufficiently confident that the other receiver has received a similar signal).

In this game, as in the two-player game, S has to decide what cost-free signal m_{θ_S} to send given her underlying quality θ_S. Therefore, let $m : \Theta_S \to \Theta_S$ represent S's strategy. Both R_1 and R_2 have to decide whether to play P or $\bar{P}$ as a function of the signal m_{θ_S}. Therefore, let $\sigma_i : \Theta_S \to \{P, \bar{P}\}$ represent R_i's strategy, with $i \in \{1, 2\}$. We restrict ourselves to pure strategies and seek Perfect Bayesian Equilibria (PBE) of the game.

Here, we wonder whether positive illusions ($m_{\theta_S} > \theta_S$) can be stabilized in this new setting. Can the uncertainty receivers face about each other's "punishment threshold" sustain positive illusions at equilibrium? Proposition 4.6.1 shows that for small enough lies, punishment can not be triggered due to R_1's uncertainty about R_2's "punishment threshold".

Proposition 4.6.1. *If $c > 0$, then at equilibrium: (i) $m^*_{\theta_S} = min\ \{\theta_S + \bar{\lambda}, \theta_R + \gamma\}$, with $\bar{\lambda} = \delta\bar{q}$, (ii) $\sigma_i(m^*_{\theta_S}) = \bar{P}$ and (iii) $\mathbb{E}_i[\theta_S | m^*_{\theta_S}] = \theta_S$, for $i \in \{1, 2\}$.*

Proposition 4.6.1 shows that under certain conditions, even though both receivers know that S is sending a lie, they will refrain from punishing S due to R_1's uncertainty about R_2's "punishment threshold". For R_1 to be willing to punish S upon observing a deceptive signal, he must believe that R_2 will punish S with a

probability greater than $\bar{q}$ (see above). Now, if the lie λ is sufficiently small ($\bar{\lambda} \leq \delta\bar{q}$), even though R_1 knows that S has sent a deceptive signal, R_1 will refrain from punishing S given that he is not confident enough that R_2 will punish too. Knowing that, even though R_2 also observes a lie, he will refrain from punishing, given that he expects R_1 not to punish. Therefore, in situations in which receivers want to coordinate their responses (e.g., punish if and only if others punish), shared knowledge about a transgression does not necessarily translate into common knowledge of a transgression, and lies can remain unpunished due to the uncertainty receivers have about each others' "punishment threshold". For big enough lies ($\bar{\lambda} > \delta\bar{q}$), both receivers will be sufficiently confident that both have observed a transgression and punishment will be triggered. This underlines the important role that *plausibility* (small enough lies that can prevent coordination) and *higher-order beliefs* (beliefs about what others believe) might have in sustaining positive illusions at equilibrium. Finally, one can see that the smaller the uncertainty (the smaller δ), the smaller the lie λ needs to be at equilibrium. Similarly, the smaller $\bar{q}$, which is the minimum level of confidence needed for the receivers to be willing to play P, the smaller the lie λ has to be.

The analysis in this section leads us to conclude that when punishment is coordinated, small lies (small positive illusions) can remain unpunished due to the uncertainty audience members have about others' "punishment threshold". This uncertainty therefore constitutes another mechanism which can stabilize positive illusions at equilibrium.

4.7 Main predictions

The models described above make several predictions. The main predictions will be listed in this section and will be confronted with empirical data in Section 4.8.

1. *Positive illusions will tend to be higher (respectively lower), the lower (respectively higher) the (ex ante) degree of observability of the underlying quality θ_S.*

   — As we have seen in the "Partner Choice" model, the degree of observability γ of the sender's underlying quality θ_S restricts the range of messages that she can send. The greater the (ex ante) observability of θ_S (the smaller γ), the smaller the range of messages that S can send. As a result, we expect positive illusions to be smaller, the greater the (ex ante) observability of θ_S.

2. *Positive illusions will tend to be higher (respectively lower), the lower (respectively higher) the (ex post) degree of observability of the underlying quality θ_S.*

   — Another prediction from the "Partner Choice" model is that positive illusions can not stabilize when the payoffs of joint projects, interactions, etc., are perfectly determined by the sender's underlying quality θ_S, that is, when θ_S is (ex post) perfectly observable. In such settings, a lie is immediately spotted by the receiver. In fact, noise in the payoffs ($z > 0$) is a necessary condition for positive illusions to persuade at equilibrium. Therefore, we expect positive illusions to be smaller, the greater the (ex post) observability of θ_S.

3. *Positive illusions will tend to be lower (respectively higher), the higher (respectively lower) the reputational costs of lying.*

   — The last prediction of our "Partner Choice" model is that the set of "Low" types that can pool with "High" types at equilibrium will be smaller, the greater the reputational costs of lying c. Therefore, we expect positive illusions to be smaller in settings in which S's reputation can suffer from being caught lying.

4. *Lies will tend to remain unpunished in contexts in which senders have sufficient plausible deniability.*

   — The main prediction of our "Community" game is that when lies (positive illusions) are small enough, then coordinated punishment by the receivers can fail to be triggered due to the inherent uncertainty about the other receivers' "punishment threshold". We therefore expect positive illusions to remain unpunished—and as a result to be stable—in contexts in which senders have sufficient plausible deniability.

4.8 Empirical evidence

This section will be dedicated at reviewing the existing literature in light of the predictions made in the above models. As discussed in Section 4.2, if positive illusions are mainly sustained by the need to promote psychological health, then we do not expect positive illusions to be affected by the observability of the trait. If, on the contrary, positive illusions are mainly sustained by information-processing or cognitive biases, then we do not expect positive illusions to be affected by the desirability

nor the observability of the trait. By contrast, the model developed in Section 4.5 suggests that if positive illusions are maintained by the effects they have on others, then the observability and desirability of the trait will be important moderators. Furthermore, the model in Section 4.6 predicts that plausible deniability can be an important factor stabilizing positive illusions. In this section, we therefore ask whether desirability, observability and plausible deniability have been observed to influence the nature of positive illusions.

4.8.1 Are positive illusions influenced by the desirability of the trait?

As described in Section 4.2.2, several researchers have argued that individual motivation is unnecessary to stabilize positive illusions. This argument is based on laboratory experiments showing that cognitive biases can account for why individuals think better about themselves than the evidence suggests. Yet, a large literature shows that trait desirability moderates the extent to which individuals self-enhance.

In a seminal paper, Weinstein (1980) has shown that individuals are more optimistically biased when it comes to desirable events. That is, individuals tend to believe that positive and desirable events are significantly more likely to happen to themselves relative to others. Similarly, Alicke (1985) has shown that the BTAE effect is larger for desirable traits (such as being kind, cooperative, bright, friendly, clever, creative or sincere), writing that "[s]ubjects perceived various traits to be more characteristic of themselves than the average college student as those traits increased in desirability" (Alicke 1985, p.1626).[7] A series of papers also describes how individuals tend to overestimate their likelihood of engaging in socially desirable behaviors, such as donating blood or contributing to charity, although they accurately estimate the propensity of others to engage in such acts (Epley & Dunning 2000, Messick et al. 1985). Moreover, Brown (2012) has shown, in a series of experiments, that the BTAE effect is significantly stronger for traits that individuals judge important for themselves (relative to unimportant traits). Finally, Tappin & McKay (2017) describe how the magnitude of positive illusions is significantly greater when judging moral qualities, such as being honest, trustworthy or fair.

Evidence suggests that the positive correlation between self-enhancement and trait desirability also appears in cross-cultural studies, particularly so in comparisons between Asians and Americans. Although some authors argue that Asians do not self-enhance (Heine et al. 1999, Heine & Hamamura 2007), others have provided

[7] This finding has been replicated by Pedregon et al. (2012) and Ziano et al. (2021).

evidence that Asians tend to self-enhance on culturally desirable traits, which are interdependent or collectivist (rather than individualistic) traits, such as loyalty (Kobayashi & Brown 2003, Sedikides et al. 2003, 2005, 2007).

Given the above-described results, why did some researchers downplay the role of individual motivation in sustaining positive illusions and rather stressed the role of non-motivational factors, such as cognitive biases? According to Brown (2012), researchers have highlighted the role of non-motivational factors especially because they have tended to study non-important traits. For instance, Kruger (1999) assessed subjects on traits such as juggling, riding a bicycle or programming a computer, while Chambers et al. (2003) asked subjects about their relative likelihood of observing a comet in the sky or being hugged by a celebrity. Similarly, Moore & Healy (2008) assess subject responses to trivia quizzes. Overall, these results suggest that the role of non-motivational factors, such as cognitive or informational biases, is most easily observed when evaluating subject on traits of little relevance to them. Moreover, in a recent meta-analysis of the BTAE literature, Zell et al. (2020) show that the BTAE effect remains robust even after controlling for the eventual role cognitive biases might play, suggesting that such biases are not sufficient to explain the stability of positive illusions.

Together, these findings suggest that the desirability of the trait is an essential moderator when it comes to the nature of positive illusions, with individuals being particularly inclined to hold enhanced views about themselves on (socially) desirable traits.

4.8.2 Are positive illusions influenced by the observability of the trait?

The studies reviewed in the previous section have stressed the important role of trait desirability, in line with the idea that individuals will self-enhance particularly on those traits which are socially desirable. In this section, we wonder whether the observability of the trait similarly moderates the nature of positive illusions.

Early findings in the literature have stressed the role of trait ambiguity as moderating the effect of the BTAE. A trait or ability is judged as ambiguous if there is room for idiosyncratic and self-serving definitions of that trait or ability (Dunning et al. 1989). This definition can be linked to the parameters γ and z in the model in Section 4.5, which described the *ex ante and ex post degree of observability* of the underlying trait. A highly ambiguous trait can be defined as a trait which is not readily verifiable since there is no clear agreed-upon metric upon which to judge that trait, whereas a non-ambiguous trait, which has an objective measure, can be

defined as a verifiable trait. Alternatively, the greater the ambiguity of the trait, the easier it is for individuals to plausibly argue that they are endowed with such a trait. As a rule, abilities, such as intelligence, are judged less ambiguous than personality traits, such as being generous or fair. As described by Allison et al. (1989):

> When a person performs a behavior that requires ability, one can infer that the person possesses the ability. When a person performs a behavior such as being fair that does not require an ability component, it is less easy to make a corresponding inference. Thus a smart behavior requires intelligence, but a moral behavior does not necessarily require morality. This tighter correspondence between intelligent acts and inferences of intelligence permits *less interpretational ambiguity* than is possible with moral judgments. (Allison et al. 1989, p.277, emphasis added)

As predicted by their framework, and in line with the model developed in Section 4.5, Allison et al. (1989) show that individuals tend to self-enhance more on moral rather than intelligent dimensions, concluding that "people do not exaggerate their positions on dimensions that are public, specific, and more or less objective" (Allison et al. 1989, p.290). This can be translated in the following way: when γ and/or z are close to zero (i.e., when the degree of observability is high), individuals tend to have rather accurate representations of their underlying quality, which is one of the main prediction of our model. This finding has been replicated by Van Lange & Sedikides (1998), who demonstrate that subjects are more likely to self-enhance on traits that are more desirable and less verifiable. Dunning et al. (1989) similarly stress the important role of trait ambiguity as moderating the extent to which individuals hold enhanced views about themselves, noting that "as traits became more ambiguous, subjects were more likely to provide a favorable comparison of themselves relative to their peers" (Dunning et al. 1989, p.1088). In line with the above-described studies, a recent meta-analysis of the BTAE literature demonstrates that the BTAE is larger for personality traits, compared to abilities (Zell et al. 2020). Finally, in a recent article aiming at downplaying the role of individual motivation in sustaining positive illusions, Logg et al. (2018) actually provide results in line with the model outlined in Section 4.5 and in line with the above-described studies. They show that when traits are given specific and precise definitions (when ambiguity low), individuals do not necessarily self-enhance. Rather, the authors find that the BTAE effect is strongest for ambiguous traits. This, again, is in line with the idea that the greater the observability of the trait, the easier it is for observers to verify the subject's claims and the less room there is for self-enhancement.

Other studies provide support for the idea that observability moderates the

nature of positive illusions in a more indirect way. For instance, Alicke et al. (1995) describe how the BTAE is reduced when individuals have to compare themselves to a specific person rather than an abstract (ambiguous) entity, such as the average college student. This result presumably stems from the fact that comparisons among pairs of individuals are more easily observable and verifiable than comparisons with an abstract entity such as the average person. Similarly, Tice et al. (1995) have shown that self-enhancement is reduced when individuals interact with their friends compared to strangers, presumably because friends have more information about the target individual's underlying traits, thereby diminishing the opportunity to self-enhance without being caught lying. Furthermore, Sedikides et al. (2002) describe how self-enhancement is curtailed when individuals are expected to explain, justify and defend their self-evaluations to another person. Having to explain and justify their self-evaluation forces individuals to provide evidence for their claims, which automatically reduces the extent to which they can self-enhance without being caught lying. Together, these results suggest that when individuals are expected to provide evidence for their claims, they tend to have accurate representations of their traits and abilities.

Finally, the literature in experimental economics studying biased updating also adds support for the moderating role of observability. In a recent paper, Drobner (2022) attempts to reconcile conflicting results coming out of the laboratory, with some authors finding that subjects tend to update their beliefs about ego-relevant information optimistically (Eil & Rao 2011), while others fail to find such optimistic updating (Coutts 2019, Ertac 2011). Drobner (2022) finds that the heterogeneity in the results arises from differences in the methodologies used in the different papers. In particular, he argues—and experimentally demonstrates—that subjects in the laboratory update their beliefs optimistically only if they fail to receive immediate feedback about their performance and if the experimenter is unaware about the realized state of the world (the subject's true performance). When subjects are told that their true performance will be revealed at the end of the experiment (to themselves and the experimenter), they fail to update optimistically (their beliefs are close to the Bayesian benchmark). This result is consistent with findings that positive illusions tend to fade when feedback draws near (Sweeny & Krizan 2013, Taylor & Shepperd 1998). These results can be interpreted in light of the model developed in Section 4.5: when subjects expect their underlying quality to be revealed to others (when the ex post observability of θ_S is high), they will tend to have rather accurate representations about their traits and abilities since there is no room for self-enhancement.

Overall, the studies reviewed in this section stress the important role of observability as moderating the nature of positive illusions, thereby providing more support to the ideas developed in Section 4.5.

4.8.3 Can plausible deniability prevent punishment?

This last section will be dedicated at investigating whether there is any empirical evidence for the claim (made in Section 4.5) that receivers punish senders when a lie is detected, and that plausible deniability might prevent punishment and hence contribute to the stability of positive illusions (a claim made in Section 4.6).

To investigate whether receivers punish senders when they learn that senders have made exaggerated claims, it is convenient to turn to the rich literature that has explored whether expressions of confidence from the part of the sender had any influence over the way in which receivers interpret a message and how receivers react after receiving feedback about the accuracy (or lack thereof) of the message. Early findings have noted the existence of a *confidence heuristic*, whereby individuals tend to trust confident senders more, supposedly inferring that their confidence must stem from greater knowledge, competence or correctness (Price & Stone 2004, Tenney et al. 2007, 2008). An alternative interpretation for why confident individuals are initially believed more—and in line with the model developed in Section 4.5—is that confidence can be seen as an expression of commitment to the claim from the part of the sender, and that by expressing confidence, the sender signals that she is ready to incur reputational costs if the claim turns out to be wrong (Vullioud et al. 2017). According to this latter argument, receivers initially trust confident individuals given that they expect confident individuals to suffer reputational costs if their claim turns out to be misguided. The findings in the above-cited papers suggest that receivers take expressions of confidence as informative signals, and not just cheap-talk, indicating that there must exist mechanisms that keep signals honest (on average).

Is there any evidence that expressions of confidence are informative due to the fact that senders tend to suffer greater costs (relative to non confident senders) when their claims appear to be misguided? Early findings have suggested that this is indeed the case. Tenney, MacCoun, Spellman & Hastie (2007, p.46) have shown that "errors in testimony damage the overall credibility of witnesses who were confident about the erroneous testimony more than that of witnesses who were not confident about it", indicating that expressions of confidence—while initially endowing the sender with greater persuasive power—lead to greater costs if the confidence appears to be unjustified. Similarly, Tenney et al. (2008) describe how those individuals who

show greater confidence in their claims tend to be trusted less when their claims turn out to be unwarranted. Further evidence comes from Vullioud et al. (2017), who presented subjects with confident and non confident senders. At first, subjects trusted the confident sender more (in line with the *confidence heuristic*), but once the confident sender's advice was revealed to be misguided, the subjects tended to adjust their trust such that the non confident sender was now trusted more. These findings suggest that receivers monitor the sender's claims together with their degree of confidence in these claims, and that once feedback indicates that the claim appears to be misguided, those senders who express the highest degree of confidence in their claims tend to suffer the greatest reputational costs. This provides support for the assumption made in the model presented in Section 4.5 that receivers are expected to impose reputational costs to senders who appear to have made overblown claims about their quality (or abilities).

Yet, other findings suggest that this conclusion might be rather premature. A collection of studies has shown that confident but wrong senders do not suffer high reputational costs, implying that lies might not always get punished (Kennedy et al. 2013, Sah et al. 2013). Can these findings be reconciled with the evidence reviewed above? A recent paper indicates that they can. Puzzled by the conflicting evidence, Tenney et al. (2019) set out to investigate whether the channel of confidence expression plays a role in determining whether overconfidence, when revealed to receivers, is punished or not. They note that papers that find overconfidence being punished by receivers have senders express their confidence verbally, while papers that fail to find such punishment have senders express their confidence nonverbally. As a consequence, Tenney et al. (2019) explore whether receivers react differently when confidence is expressed verbally or nonverbally. Their argument, which they call the *plausible deniability hypothesis*, is that "nonverbal modes of confidence expression provide plausible deniability because, typically, these expressions are not as clearly tied to specific levels of confidence about performance as are verbal expressions" (Tenney et al. 2019, p.398). In other words, the argument is that verbal expressions of confidence are harder-to-deny than nonverbal expressions of overconfidence. Their prediction is therefore that senders who express confidence verbally will be more likely to be punished, if receivers obtain evidence that the sender's claim is unwarranted, compared to senders who express their confidence nonverbally, exactly because the latter have more room to plausibly deny that they have self-enhanced. Through a series of experiments, Tenney et al. (2019) find support for their hypothesis, noting, among other things, that those senders who express confidence verbally incur higher reputational costs (compared to senders who ex-

press their confidence nonverbally) when their performance is shown to be lower than their stated level of confidence, and that a denial of overconfidence from the part of senders who expresses overconfidence nonverbally is judged more plausible by the receivers (compared to senders who deny having expressed overconfidence verbally).

These findings therefore indicate that plausible deniability can prevent punishment. Although the authors suggest that plausible deniability prevents punishment because of psychological factors (such as memory or dispositions towards the sender), the model developed in Section 4.6 predicts that plausible deniability can prevent punishment because it potentially hampers coordinated punishment among receivers. The argument is that when senders can plausibly deny that they self-enhanced (or that they have been *over*confident), this creates uncertainty among receivers which can prevent coordinated action—even though each receiver might be inclined to punish. This argument is analogous to what has been described as the *omission strategy*. According to DeScioli et al. (2011, p.445), "people choose omissions to avoid third-party condemnation and punishment", given that omissions have been shown to be less likely to be punished than commissions (Spranca et al. 1991), presumably because omissions do not generate the same degree of common knowledge among observers than commissions do (Hoffman et al. 2018). The idea is therefore that individuals judge omissions to be "less bad" than commissions, with their judgment presumably reflecting equilibrium play. I argue that the same mechanism explains why overconfidence remains unpunished—and judged less harshly—when expressed nonverbally.

4.9 Conclusion

This chapter has been concerned with the origins and stability of positive illusions, defined as overly positive beliefs about oneself. In a two-player "Partner Choice" setting, I have investigated whether a Sender (S) can persuade a Receiver (R) to accept her as a partner by adopting positive illusions. That is, the focus has been on whether positive illusions can effectively persuade at equilibrium. In a three-player "Community" game, I have investigated whether positive illusions can remain stable in a setting in which a Sender, whose objective is to appear as better than she is, faces two Receivers (an audience) who have to decide whether or not to punish S after observing the results of a task undertaken by S. Given that Receivers want to coordinate their response, the focus of this model has been on whether small enough lies—which might confer S enough plausible deniability—can prevent coordinated

punishment.

The results of the "Partner Choice" game show that when payoffs of the interaction are deterministic, then positive illusions can not persuade, given that lies are immediately spotted by R. On the other hand, when payoffs are non-deterministic (when R can not infer with certainty S's underlying quality from the payoffs of the interaction), "Low" types can pool with "High" types by adopting positive illusions at equilibrium. Adopting positive illusions can therefore persuade R to accept S, even though R is not fooled at equilibrium and correctly anticipates the average value of S's type. Importantly, the analysis predicts that positive illusions will be sensitive to the reputational costs of lying and to the (*ex ante* and *ex post*) degree of observability of S's underlying quality. The results of the "Community" game confirm that small enough lies can prevent coordinated punishment of S by the Receivers. That is, in settings in which Receivers want to punish a self-enhancing S if and only if they expect other Receivers to punish too, "small lies" (small illusions) can generate enough uncertainty among the Receivers about each other's "punishment threshold", which ultimately prevents coordination. The prediction is that the size of the illusion will be sensitive to the ease with which Senders can plausibly deny having lied.

The empirical literature on positive illusions appears to be in line with the predictions of the "Partner Choice" and the "Community" games, in the sense that positive illusions seem to be sensitive to the desirability and the degree of observability of the trait, as well as to the reputational costs of lying. Moreover, there is some evidence that Senders remain unpunished when they have enough plausible deniability about having self-enhanced. Interestingly, alternative theories about the origins of positive illusions do not make such predictions. This suggests that positive illusions may have a persuasive function, and may thereby be sustained by their interpersonal effects.

4.10 Appendix

Proof of Proposition 4.5.1. (i) Assume that $\mathbb{E}[\theta_S|m_{\theta_S}] = m_{\theta_S}$, and assume, WLOG, that $\theta_S < \theta_R$. If S sends $m_{\theta_S} \in (\theta_R, \theta_R + \gamma]$, then R accepts to interact with S. S would derive benefits equal to $\beta\theta_R > 0$ from her deception. Now, if payoffs are deterministic, then when $\pi_R(\theta_S, \theta_R)$ (strictly inferior to $\mathbb{E}[\pi_R(m_{\theta_S}, \theta_R)]$) realizes, R infers that S's underlying quality must necessarily be $\theta_S < m_{\theta_S}$. S therefore incurs cost c. It follows that as long as $c \geq \beta\theta_R$, S will not find it profitable to send enhanced signals about her quality. In fact, if $c \geq \beta\theta_R$, then there exists a

separating equilibrium at which $m^*_{\theta_S} = \theta_S$ for all $\theta_S \in \Theta_S$, and R accepts S if and only if $m^*_{\theta_S} > \theta_R$. To see why this is an equilibrium, consider first R's strategy. If S truthfully reveals her underlying quality θ_S at equilibrium, then R's equilibrium expectation is $\mathbb{E}^*[\theta_S|m^*_{\theta_S}] = \theta_S$ and R's payoff-maximizing strategy is to accept any S such that $m^*_{\theta_S} > \theta_R$. Now, consider a deviation m'_{θ_S} for S with type θ_S. Given R's equilibrium strategy, it can never be profitable for S to send $m'_{\theta_S} < \theta_S$. Sending $m'_{\theta_S} > \theta_S$ might only be profitable if $m'_{\theta_S} > \theta_R$, and $\theta_S < \theta_R$. Nevertheless, if S with underlying quality $\theta_S < \theta_R$ sends $m'_{\theta_S} > \theta_S > \theta_R$, then R accepts S (given his equilibrium beliefs), providing benefits $\beta\theta_R > 0$ to S but also generating costs c due to the lie being spotted. Therefore, as long as $c \geq \beta\theta_R$, sending $m'_{\theta_S} > \theta_S$ can never be profitable for S. It follows that when $c \geq \beta\theta_R$, there exists a fully truthful equilibrium.

(ii) If $0 \leq c < \beta\theta_R$, then R knows that he can not prevent S from lying about θ_S by threatening to damage her reputation. R must then expect that S will send an enhanced signal about her underlying quality θ_S ($\theta_R \leq m_{\theta_S} \leq \theta_R + \gamma$) in order to try to convince R to accept her. The signal m_{θ_S} therefore becomes uninformative. In fact, there exists a *babbling* equilibrium at which S sends $m^*_{\theta_S} = \theta_R + \gamma$, for all $\theta_S \in \Theta_S$ and R always denies S. To see that this is an equilibrium, consider first R's equilibrium beliefs. Given S's equilibrium strategy, R relies only on his prior when deciding whether to accept or deny S. Since R's prior is uniformly distributed over Θ_S, then upon receiving S's signal $m^*_{\theta_S} = \theta_R + \gamma$, R's equilibrium expectation $\mathbb{E}^*[\theta_S|m^*_{\theta_S}]$ has to be equal to θ_R. R's expected payoff is therefore $\mathbb{E}_{\theta_S}[\pi_R] = \phi_R + \alpha(\theta_R - \theta_R) = \phi_R$, which is equal to his outside option. R therefore decides to deny S (see Footnote 2). If R receives a signal $m'_{\theta_S} < \theta_R + \gamma$, then we can assume that $\mathbb{E}^*[\theta_S|m'_{\theta_S}] = \theta_R - \gamma$. Given R's equilibrium beliefs, S is not incentivized to deviate from her equilibrium strategy $m^*_{\theta_S} = \theta_R + \gamma$, for all $\theta_S \in \Theta_S$. Again, a deviation $m'_{\theta_S} < \theta_R + \gamma$ can not be profitable for S with type θ_S, while a deviation $m'_{\theta_S} > \theta_R + \gamma$ is not feasible. Therefore, when $0 \leq c < \beta\theta_R$, the signal $m^*_{\theta_S}$ is uninformative at equilibrium. ■

Proof of Proposition 4.5.2. At any semi-separating equilibria with cutoff value $\theta^*_S \in [\theta_R - \gamma, \theta_R + \gamma]$, such that all types $\theta_S < \theta^*_S$ send $m^{l^*}_{\theta_S} = \frac{\theta^*_S + (\theta_R - \gamma)}{2} =$ "Low" and are denied by R, and all types $\theta_S \geq \theta^*_S$ send $m^{h^*}_{\theta_S} = \frac{\theta^*_S + (\theta_R + \gamma)}{2} =$ "$High$" and are accepted by R, R's equilibrium expectation must be $\mathbb{E}^*[\theta_S|m^{l^*}_{\theta_S}] = \frac{\theta^*_S + (\theta_R - \gamma)}{2}$ when receiving the signal $m^{l^*}_{\theta_S}$, and $\mathbb{E}^*[\theta_S|m^{h^*}_{\theta_S}] = \frac{\theta^*_S + (\theta_R + \gamma)}{2}$ when receiving the signal $m^{h^*}_{\theta_S}$. It follows that if the cutoff type θ^*_S decides to send an enhanced signal about her underlying quality, then the probability that she is caught lying is equal to $\frac{\alpha[\theta_R + \gamma - \theta^*_S]}{4z}$. This is due to the fact that R's payoff when interacting with the cutoff type θ^*_S will

necessarily fall inside the interval $I = [\phi_R + \alpha(\theta_S^* - \theta_R) - z, \phi_R + \alpha(\theta_S^* - \theta_R) + z]$, but if the cutoff type θ_S^* sends $m_{\theta_S}^{h^*} = \frac{\theta_S^* + (\theta_R + \gamma)}{2} = "High"$, then R will *expect* to observe a payoff in the interval $I' = [\phi_R + \alpha(\frac{\theta_S^* + (\theta_R + \gamma)}{2} - \theta_R) - z, \phi_R + \alpha(\frac{\theta_S^* + (\theta_R + \gamma)}{2} - \theta_R) + z]$. The probability that the realized payoff in I will be lower than the lower bound of I' is equal to $\frac{[\alpha(\frac{\theta_R + \gamma + \theta_S^*}{2}) - \alpha\theta_S^*]}{2z} = \frac{\alpha[\theta_R + \gamma - \theta_S^*]}{4z}$.

At equilibrium, the cutoff type θ_S^* must be indifferent between sending $m_{\theta_S}^{l^*}$ and be denied by R, and sending $m_{\theta_S}^{h^*}$ and be accepted by R. The following equality must therefore necessarily hold:

$$\phi_S = \frac{\alpha[\theta_R + \gamma - \theta_S^*]}{4z}[\phi_S + \beta\theta_R - c] + [1 - \frac{\alpha[\theta_R + \gamma - \theta_S^*]}{4z}][\phi_S + \beta\theta_R]$$
$$\theta_S^* = (\theta_R + \gamma) - \frac{4z\beta\theta_R}{\alpha c}.$$

■

Proof of Proposition 4.5.3. In order to have $\theta_S^* \in (\theta_R - \gamma, \theta_R)$ at equilibrium, we need:

1.

$$\begin{aligned} \theta_S^* &> \theta_R - \gamma \\ (\theta_R + \gamma) - \frac{4z\beta\theta_R}{\alpha c} &> \theta_R - \gamma \\ 2\gamma &> \frac{4\theta_R z\beta}{\alpha c} \\ c &> \frac{2\theta_R z\beta}{\alpha\gamma}. \end{aligned}$$

2.

$$\begin{aligned} \theta_S^* &< \theta_R \\ (\theta_R + \gamma) - \frac{4z\beta\theta_R}{\alpha c} &< \theta_R \\ \gamma &< \frac{4\theta_R z\beta}{\alpha c} \\ c &< \frac{4\theta_R z\beta}{\alpha\gamma}. \end{aligned}$$

■

Proof of Proposition 4.5.4. We know from Proposition 4.5.2 and Proposition 4.5.3 that at any semi-separating equilibrium with cutoff value $\theta_S^* \in [\theta_R - \gamma, \theta_R + \gamma]$, such

that all types $\theta_S < \theta_S^*$ send $m_{\theta_S}^{l^*} = \frac{\theta_S^* + (\theta_R - \gamma)}{2} =$ "Low" and are denied by R, and all types $\theta_S \geq \theta_S^*$ send $m_{\theta_S}^{h^*} = \frac{\theta_S^* + (\theta_R + \gamma)}{2} =$ "$High$" and are accepted by R, the cutoff type is given by $\theta_S^* = (\theta_R + \gamma) - \frac{4z\beta\theta_R}{\alpha c}$, and we need $\frac{2\theta_R z\beta}{\alpha\gamma} < c < \frac{4\theta_R z\beta}{\alpha\gamma}$ to hold for $\theta_S^* \in (\theta_R - \gamma, \theta_R)$ to be satisfied.

For all θ_S, the extra gain of choosing to send $m_{\theta_S}^{h^*}$ instead of $m_{\theta_S}^{l^*}$ is equal to $\Delta_{\pi_S}(\theta_S) = -c[\frac{\alpha[\theta_R + \gamma - \theta_S]}{4z}] + \beta\theta_R$. The derivative of $\Delta_{\pi_S}(\theta_S)$ with respect to θ_S is positive, which implies that if type θ_S prefers to send $m_{\theta_S}^{h^*}$, then all types $\theta_S^{'} > \theta_S$ prefer to send $m_{\theta_S}^{h^*}$. Therefore, if type θ_S^* is indifferent between sending $m_{\theta_S}^{l^*}$ and $m_{\theta_S}^{h^*}$, then all types $\theta_S < \theta_S^*$ are incentivized to send $m_{\theta_S}^{l^*}$ at equilibrium, and all types $\theta_S > \theta_S^*$ are incentivized to send $m_{\theta_S}^{h^*}$. Moreover, R is incentivized to deny S when receiving $m_{\theta_S}^{l^*}$, given that his expected payoff is equal to $\mathbb{E}_{\theta_S}[\pi_R] = \phi_R + \alpha(\frac{\theta_S^* + (\theta_R - \gamma)}{2} - \theta_R) < \phi_R$, while R is incentivized to accept S when receiving $m_{\theta_S}^{h^*}$, given that his expected payoff is equal to $\mathbb{E}_{\theta_S}[\pi_R] = \phi_R + \alpha(\frac{\theta_S^* + (\theta_R + \gamma)}{2} - \theta_R) > \phi_R$.

If R receives a signal $m_{\theta_S}^{'} \neq m_{\theta_S}^{h^*}$ or $m_{\theta_S}^{'} \neq m_{\theta_S}^{l^*}$, then we can simply assume that $\mathbb{E}^*[\theta_S | m_{\theta_S}^{'}] = \theta_R - \gamma$.

■

Proof of Proposition 4.6.1. Given that we have assumed that the only uncertainty lies in R_1's beliefs about R_2's beliefs about π_{θ_S}, all that matters for the purpose of this proof is to understand the conditions under which R_1 is willing to play P or $\bar{P}$. We know that R_1 will be willing to play P as long as he expects R_2 to play P with probability (strictly) greater than $\bar{q}$.[8] This implies that upon observing a lie (i.e., $m_{\theta_S} = \theta_S + \lambda > \Pi_{R_1} = \theta_S$), R_1 must believe that R_2 has observed a lie with probability greater than $\bar{q}$. In fact, for R_1 to be sufficiently confident that R_2 will spot $m_{\theta_S} = \theta_S + \lambda$ as lie, it needs to be the case that the proportion of values of $z \in [0, \delta]$ such that if R_2 were endowed with such a value of z, he would be willing to play P upon observing $m_{\theta_S} = \theta_S + \lambda$, needs to be greater than $\bar{q}$. Upon receiving $m_{\theta_S} = \theta_S + \lambda$, the proportion of values of z such that if R_2 were endowed with such a value of z, he would be willing to play P, is written $\frac{(\theta_S + \lambda) - \theta_S}{(\theta_S + \delta) - \theta_S} = \frac{\lambda}{\delta}$. For R_1 to be willing to play P, it therefore needs to be the case that $\frac{\lambda}{\delta} > \bar{q}$, or that $\lambda > \delta\bar{q}$. This implies that for all values of λ strictly greater than $\delta\bar{q}$, R_1 would be sufficiently confident that R_2 has spotted $m_{\theta_S} = \theta_S + \lambda$ as a lie too, and R_1 would therefore be willing to play P. Expecting this, R_2 is therefore willing to play P if and only if $\lambda > \delta\bar{q}$. It follows that as long as S sends a signal $m_{\theta_S} = \theta_S + \lambda$, with $\lambda \leq \delta\bar{q}$, R_1 will not be sufficiently confident that R_2 considers this as a lie, and will therefore refrain from playing P. Expecting this, R_2 also refrains from playing P when $\lambda \leq \delta\bar{q}$. S's

[8] The assumption that R_1 is willing to play P only if the probability that R_2 plays P is *strictly* greater than $\bar{q}$ is made to ensure the existence of an equilibrium.

payoff maximizing strategy is therefore to send $m^*_{\theta_S} = \theta_S + \bar{\lambda}$, with $\bar{\lambda} = \delta\bar{q}$, given that $\bar{\lambda}$ is the greatest lie that goes unpunished. By playing $m^*_{\theta_S} = \theta_S + \bar{\lambda}$, S receives payoffs equal to $\theta_S + \bar{\lambda}$. Sending a signal $m'_{\theta_S} > \theta_S + \bar{\lambda}$ would trigger coordinated punishment from the part of the receivers, and S would incur a cost c. As long as $c > 0$, S would want to refrain from sending $m'_{\theta_S} > \theta_S + \bar{\lambda}$. Therefore, if $\gamma > 0$ and $c > 0$, S will send $m^*_{\theta_S} = \theta_S + \bar{\lambda}$, with $\bar{\lambda} = \delta\bar{q}$. Now, given that both R_1 and R_2 observe $m^*_{\theta_S} = \theta_S + \bar{\lambda}$, they both *know* that S has sent a lie. Moreover, since they both know the (deterministic) payoff function, their equilibrium expectation will be $\mathbb{E}_1[\theta_S|m^*_{\theta_S}] = \mathbb{E}_2[\theta_S|m^*_{\theta_S}] = \theta_S$. It follows that even though both receivers know that S has lied about her underlying quality θ_S, they will refrain from playing P due to R_1's uncertainty about R_2's willingness to punish S.

What happens if s sends a different message at equilibrium? If $\theta_S + \bar{\lambda} < \theta_R + \gamma$, with $\bar{\lambda} = \delta\bar{q}$, then:

— If s sends $m'_{\theta_S} > \theta_S + \bar{\lambda} \leq \theta_R + \gamma$, then $\sigma_i(m'_{\theta_S}) = P$ and $\mathbb{E}_i[\theta_S|m'_{\theta_S}] = \theta_S$, for $i \in \{1, 2\}$ and s can increase her equilibrium payoff by sending $m^*_{\theta_S} = \theta_S + \bar{\lambda}$.

— If s sends $m''_{\theta_S} < \theta_S + \bar{\lambda} \geq \theta_R - \gamma$, then $\sigma_i(m''_{\theta_S}) = \bar{P}$ and $\mathbb{E}_i[\theta_S|m''_{\theta_S}] = \theta_S$, for $i \in \{1, 2\}$ and s can increase her equilibrium payoff by sending $m^*_{\theta_S} = \theta_S + \bar{\lambda}$.

On the other hand, if $\theta_S + \bar{\lambda} \geq \theta_R + \gamma$, with $\bar{\lambda} = \delta\bar{q}$, then:

— s can not send $m'_{\theta_S} > \theta_R + \gamma$.

— If s sends $m''_{\theta_S} < \theta_R + \gamma \geq \theta_R - \gamma$, then $\sigma_i(m''_{\theta_S}) = \bar{P}$ and $\mathbb{E}_i[\theta_S|m''_{\theta_S}] = \theta_S$, for $i \in \{1, 2\}$ and s can increase her equilibrium payoff by sending $m^*_{\theta_S} = \theta_R + \gamma$.

■

Chapter 5

Distance in Beliefs and Individually-Consistent Sequential Equilibrium[1]

(co-authored with Gisèle Umbhauer)

Summary

The concept of Individually-Consistent Sequential-Equilibrium broadens the concept of Sequential Equilibrium by allowing players to have different beliefs on potential deviations. This heterogeneity spontaneously gives rise to a notion of distance between beliefs. Yet, studying the distance between beliefs in a strategic context reveals to be intricate. Announced beliefs may be different from revealed beliefs and the meaning of distance depends on the role assigned to beliefs. If out-of-equilibrium beliefs help getting a larger payoff at equilibrium, then we might need to reconsider the traditional definition of sequential rationality: more than just requiring that players behave optimally at every information set given their beliefs and the strategies played by other players, we might additionally require that there does not exist another perturbation scheme that is individually-consistent and which provides higher payoffs to the players.

[1] Another version of this chapter has appeared as BETA Working Paper: Umbhauer, G. & Wolff, A. (2019), 'Individually-Consistent Sequential Equilibrium', *BETA Working Paper*, N° 2019-39.

Classification

JEL Classification: C72

Keywords: AGM-Consistency, Distance in Beliefs, Heterogeneous Beliefs, Individually-Consistent Sequential Equilibrium, Revealed Beliefs

5.1 Introduction

In this chapter, we extend the concept of Individually-Consistent Sequential Equilibrium (ICSE, Umbhauer & Wolff 2019), which builds on the Sequential Equilibrium (SE, Kreps & Wilson 1982), a solution concept commonly used to solve extensive-form games. The SE requires consistency of beliefs at all information sets, even at those that find themselves out of the equilibrium strategy path. This implies that players are required to share the same beliefs at out-of-equilibrium information sets, even about the numerical values of mathematical artifacts used to generate perturbations of strategy profiles, which are arbitrary by nature. Since there is no *a priori* basis for requiring players to agree on the probabilities of other players' possible mistakes (or deviations), the ICSE accepts different perturbation systems for different players.

This chapter focuses on games with $n \geq 3$ players since these are the games in which the ICSE solution concept can differ from the SE. In particular, we focus on games in which some players might belong to a same social group or a same community. Therefore, although out-of-equilibrium beliefs are never directly confronted to reality (so that players can in some sense *agree to disagree*), players that belong to a same social group may feel ill at ease when adopting different beliefs. In fact, research in political science has shown that individuals are often motivated to shift their beliefs towards the ones associated with the social groups they belong to (Barber & Pope 2019, Gould & Klor 2019, Slothuus & Bisgaard 2021). Therefore, rather than being completely arbitrary, beliefs at out-of-equilibrium information sets might be correlated among players. This leads us to develop a notion of distance between the beliefs of different players as well as the idea of maximally allowed heterogeneity between the players' beliefs.

The notion of distance between beliefs introduced in this chapter can not be properly studied without further delving into the function of beliefs and their intrinsic link to actions. For instance, a player may publicly declare to hold some beliefs but his actions contradict the proclaimed beliefs. That is to say, the *revealed* beliefs of the player are different from the *announced* beliefs. The question then arises as to whether we should measure the distance between the player's announced beliefs or between their revealed beliefs. Furthermore, in a game as in real life, the purpose of out-of-equilibrium beliefs may be to help a player maximize his own payoffs. In this sense, beliefs become strategic and they may in some way belong to the strategy set of the players. For instance, it might be in Player 1 and Player 2's interests to have similar (or different) beliefs so as to incentivize Player 3 to adopt a strategy

that maximizes their own payoffs. In this context, players build their beliefs with a strategic purpose and the distance between their beliefs becomes irrelevant. These considerations lead us to revisit the notion of sequential rationality in dynamic games of incomplete information. More than just requiring that players behave optimally at every information set given their beliefs and the strategies played by other players, we might additionally require that there does not exist another perturbation scheme that is individually-consistent and which provides higher payoffs to the players.

The chapter is organized as follows. In Section 2, we describe the concept of ICSE and discuss the way it introduces heterogeneity in beliefs at out-of-equilibrium information sets. In Section 3, we compare our concept with other often-used solution concepts such as PBE or AGM-consistency. In Section 4, we turn to the notion of distance between beliefs. We present two ways of measuring distance. The first is an order relation on beliefs, while the second is an Euclidean notion of distance, in that we measure the minimal payoff perturbations necessary to ensure convergence in beliefs. Yet, the main difficulties remain elsewhere. In Section 5, we introduce the distinction between revealed and announced beliefs and discuss the strategic function of beliefs. Section 6 discusses the findings of this chapter by revisiting the definition of sequential rationality. The last section concludes.

5.2 Individually-Consistent Sequential Equilibrium

In this chapter, we consider finite extensive-form games and focus on games with $n \geq 3$ players. Let N represent the finite set of players (with typical element $n \in N$), X the set of non-terminal decision nodes (with typical element $x \in X$) and H the set of all possible information sets (with $h \in H$ a specific information set). Let $H_i \subseteq H$ denote the set of all possible information sets at which Player i might be called upon to play. We call $i(h)$ the player playing at h and for every $h \in H_i$, we note A_h the set of actions available to player i at information set h.

A *behavioral strategy* for player i, noted π_i, is a probability distribution over her possible actions at each of her information sets. That is, a behavioral strategy for player i is a member of $\bigtimes_{h \in H_i} \Delta(A_h)$. The set of behavioral-strategy profiles is therefore $\bigtimes_{i \in N} \bigtimes_{h \in H_i} \Delta(A_h)$, with typical element $\pi = (\pi_i)_{i \in N}$. A *system of beliefs* is a function $\mu : X \to [0, 1]$ such that $\forall h \in H, \Sigma_{x \in h} \mu(x) = 1$.

The Sequential Equilibrium (SE) requires *consistency* of beliefs at all information sets, even at those that find themselves out of the equilibrium strategy path. To generate beliefs that are consistent at every information set, Kreps & Wilson (1982)

require that the belief vector μ be the limit of a sequence of belief vectors derived from Bayes' rule applied to a sequence of *fully mixed* strategy profiles (strategy profiles that put positive probability to every action in every information set). Let us denote by $\times_{h\in H}\Delta^0(A_h)$ the set of all fully mixed behavioral strategies. Formally, a pair (μ,π) is consistent if and only if there exists some sequence $(\hat{\mu}^k,\hat{\pi}^k)_{k=1}^{\infty}$ such that:

1. $\hat{\pi}^k \in \times_{h\in H}\Delta^0(A_h), \forall k \in \{1,2,3,...\}$,
2. $\hat{\mu}_h^k(x) = \frac{P(x|\hat{\pi}^k)}{\sum\limits_{y\in h} P(y|\hat{\pi}^k)}, \forall h \in H, \forall x \in h, \forall k \in \{1,2,3,...\}$,[2]
3. $\pi_{i(h)}(a_h) = \lim\limits_{k\to\infty} \hat{\pi}^k(a_h), \forall i \in N, \forall h \in H, \forall a_h \in A_h$,
4. $\mu_h(x) = \lim\limits_{k\to\infty} \hat{\mu}_h^k(x), \forall h \in H, \forall x \in h$.

A SE is defined to be any pair (μ,π) that is consistent and sequentially rational (Kreps & Wilson 1982, p.872).

What is crucial in the concept of SE is that the players are required to implicitly agree on the value of the ϵ used to generate perturbations of the strategy profiles. That is, while the ϵ are arbitrary in nature (they only represent mathematical artifacts), players still need to share the same beliefs about their numerical values. In some way, it is as if an external player shakes the strategies for everybody. In Umbhauer & Wolff (2019), we argue that this requirement is too strong. Indeed, we argue that there is no *a priori* basis for requiring that players agree on the probabilities of other players' possible mistakes (or deviations).

Formally, what distinguishes our Individually-Consistent Sequential Equilibrium (ICSE) concept from the SE is that we do not require the existence of only one sequence of perturbed strategy profiles on which all players need to agree but allow for different perturbation systems for different players. In other words, each player j introduces his own perturbations on the actions at each information set $h \in H$. So, $\hat{\pi}_{j,i(h)}^k(a_h)$ is the value player j assigns to the probability with which player $i(h)$ plays a_h at his information set h, while $\hat{\pi}_j^k$ is player j's profile of perturbed strategies in the whole game. Of course, for consistency, we require that $\pi_{i(h)}(a_h) = \lim\limits_{k\to\infty} \hat{\pi}_{j,i(h)}^k(a_h)$ for all $j \in N$, so that each player j's perturbed strategy profile has to fit with the played actions in the game.

Therefore, a pair (μ,π) is *individually-consistent* if and only if there exist some sequences $(\hat{\mu}_j^k,\hat{\pi}_j^k)_{k=1}^{\infty}$, for all $j \in N$, such that:

1. $\hat{\pi}_{j,i(h)}^k \in \times_{h\in H}\Delta^0(A_h), \forall k \in \{1,2,3,...\}, \forall j \in N$,

[2] With $P(x|\cdot)$ being computed using Bayes' rule.

2. $\hat{\mu}^k_{i(h)}(x) = \frac{P(x|\hat{\pi}^k_{i(h)})}{\sum_{y\in h} P(y|\hat{\pi}^k_{i(h)})}, \forall h \in H, \forall x \in h, \forall k \in \{1, 2, 3, ...\}$,

3. $\pi_{i(h)}(a_h) = \lim_{k\to\infty} \hat{\pi}^k_{j,i(h)}(a_h), \forall h \in H, \forall a_h \in A_h, \forall j \in N$,

4. $\mu_{i(h)}(x) = \lim_{k\to\infty} \hat{\mu}^k_{i(h)}(x), \forall h \in H, \forall x \in h$.

An *Individually-Consistent Sequential Equilibrium* (ICSE) is any pair (μ, π) that is both individually-consistent and sequentially rational.

Let us illustrate the consequences of such a concept. In the game in Figure 5.1, there does not exist any SE leading Player 1 to play C_1 (see Appendix 1), so the players can not reach the Pareto optimal payoffs (5.99, 10, 10). As a matter of fact, to sustain B_2, Player 2 has to believe that Player 1 trembles toward B_1 at least 4 times more often than toward A_1 ($\mu(x_2) \leq \frac{1}{5}$), whereas to be willing to play B_3, Player 3 has to believe that Player 1 trembles toward B_1 at most 3 times more often than toward A_2 ($\mu(y_2) \geq \frac{1}{4}$, hence $\mu(x_2) \geq \frac{1}{4}$). This is not possible in a SE, in that all the players shake the strategies in the same way. Yet this becomes possible with the ICSE.

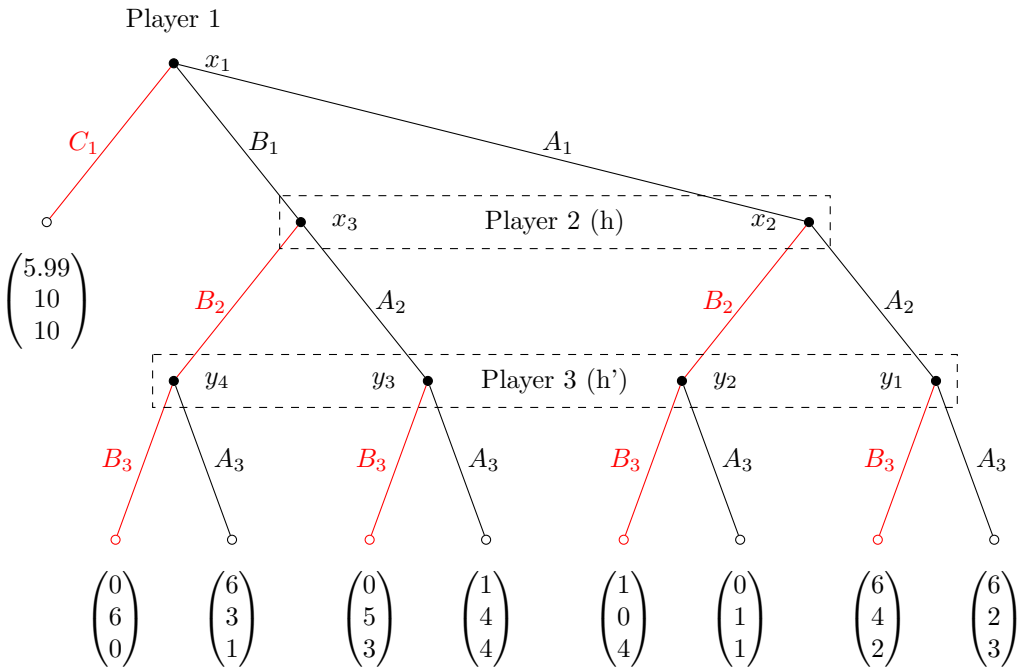


Figure 5.1 – An example of the distinction between ICSE and SE.

What is new, in comparison with the SE, is the fact that players can have different beliefs at the same out-of-equilibrium information set. So, in the above example, we can set: $\mu_2(x_2) = 0.1$, $\mu_2(x_3) = 0.9$ for Player 2, and $\mu_3(y_1) = 0$, $\mu_3(y_2) = 0.3$, $\mu_3(y_3) = 0$ and $\mu_3(y_4) = 0.7$ for Player 3. This implicitly means that

Player 3 assigns the belief 0.3 to x_2 and the belief 0.7 to x_3, given that Bayes' Rule requires that Player 3's beliefs are consistent: $\mu(y_2) = \mu(x_2)$ and $\mu(y_4) = \mu(x_3)$. This is due to the fact that each player shakes the strategies in the way he wants. So, for example, Player 2 may have in mind the perturbed strategy profile $\{(1-\epsilon^k-9\epsilon^k)C_1+9\epsilon^k B_1+\epsilon^k A_1, (1-\epsilon^k)B_2+\epsilon^k A_2, (1-\epsilon^k)B_3+\epsilon^k A_3\}$, while Player 3 may have in mind $\{(1-3\epsilon^k-7\epsilon^k)C_1+7\epsilon^k B_1+3\epsilon^k A_1, (1-\epsilon^k)B_2+\epsilon^k A_2, (1-\epsilon^k)B_3+\epsilon^k A_3\}$.[3] Heterogeneous beliefs at out-of-equilibrium information sets can therefore sustain the Pareto optimal payoffs at equilibrium in this strategic context.

5.3 Connections between ICSE and other solution concepts

5.3.1 Links between ICSE, SE, SPNE, PBE and SCE

Few solution concepts support the idea that people may share different beliefs at out-of-equilibrium information sets.[4] In fact, researchers tend to require that "players can not agree to disagree", the logical result stemming from Aumann (1976)'s paper. Yet, in our context, there is no true state to discover since the beliefs are about deviations that will never occur at equilibrium. So Player 2 (respectively Player 3) can durably think that, if they should face a deviation, then surely Player 1 played A_1 with a probability lower than 1/5 (respectively with a probability larger than 1/4). Nothing will contradict their beliefs given that Player 1 never deviates.

In the following Proposition, we enumerate the existing links between our ICSE solution concept and other well-known and often-used solution concepts, such as Subgame-Perfect Nash Equilibrium (SPNE), Sequential Equilibrium (SE), Perfect Bayesian Equilibrium (PBE) and Self-Confirming Equilibrium (SCE).

Proposition 5.3.1. *1. The set of ICSE is included in the set of Subgame-Perfect Nash Equilibria (SPNE).*

2. By construction, the set of SE is included in the set of ICSE, so the existence of an ICSE in a finite extensive-form game follows from the existence of a SE in a finite extensive-form game.

3. The set of ICSE is equal to the set of SE in a two-player game.

[3] Player 1's perturbations have no impact on the game, so we can suppose that he has the same profile of perturbed strategies as Player 2 for example.

[4] We thank Giacomo Bonanno for informing us that Greenberg et al. (2009) developed a similar idea in their MACA concept.

4. *There is no inclusion relation between the set of Perfect Bayesian Equilibria (Fudenberg & Tirole 1991) and the set of ICSE.*

5. *The set of ICSE is included in the set of Self-Confirming Equilibria (Fudenberg & Levine 1993).*

Proof. In Appendix 2. ■

5.3.2 Links between ICSE and AGM-consistency

The ICSE also shares some links with the concept of AGM-consistency (Bonanno 2013, 2016). AGM-consistency introduces a plausibility order on stories of actions and belief revision is based on this plausibility. This plausibility concept grants a large degree of freedom to the way beliefs are computed after deviations; this liberty differs from heterogeneous perturbation systems but it shares a partial link with the ICSE. The following Proposition describes these links.

Proposition 5.3.2. *i) The set of ICSE beliefs is almost included in the set of AGM-consistent beliefs.*

ii) There is no inclusion relationship between the set of ICSE beliefs and Bonanno's Perfect Bayesian Equilibrium beliefs.

Proof. In Appendix 2. ■

We come back to Bonanno's concept when studying the notion of distance, so we illustrate this concept on the game in Figure 5.1. Consider the ICSE with $\mu_2(x_2) = 0.1$, $\mu_2(x_3) = 0.9$, $\mu_3(y_1) = 0$, $\mu_3(y_2) = 0.3$, $\mu_3(y_3) = 0$ and $\mu_3(y_4) = 0.7$. These probabilities are compatible with AGM-consistency since they give positive weight to the actions (stories) A_1 and B_1 and to the stories A_1B_2 and B_1B_2. Therefore, they respect the plausibility-preserving action B_2. AGM-consistency is a qualitative notion, so the values of the beliefs are not important. What matters is that if the support of the beliefs are the stories A_1 and B_1, then the support of the stories reaching h' are the stories A_1B_2 and B_1B_2. So, every ICSE, which by definition lead to $\mu_2(x_2) \leq \frac{1}{5}$, $\mu_2(x_3) = 1 - \mu_2(x_2)$, and $\mu_3(y_1) = 0$, $\mu_3(y_2) \geq \frac{1}{4}$, $\mu_3(y_3) = 0$, $\mu_3(y_4) = 1 - \mu_3(y_2)$ respect AGM-consistency, except for the assessment that puts a 0 on $\mu_(x_2)$ or a 1 on $\mu_(y_2)$.

As a matter of fact, let us consider the "extreme" ICSE with $\mu_2(x_2) = 0$, $\mu_2(x_3) = 1$, $\mu_3(y_1) = 0$, $\mu_3(y_2) = 1$, $\mu_3(y_3) = 0$ and $\mu_3(y_4) = 0$. According to AGM-consistency, plausible histories can not sustain these beliefs, given that if $\mu(B_1)$ (the probability assigned to story B_1) is equal to 1 (because $\mu(x_2) = 0$ and

$\mu(x_3) = 1$), then $\mu(B_1B_2)$ (the probability assigned to story B_1B_2) is also 1, since B_2 is played with probability 1 (it is the plausibility-preserving action); so the story B_1B_2 is as plausible as the story B_1. Given that A_1 is a less plausible story (in fact $\mu(A_1) = 0$) and given that $\mu(A_1) = \mu(A_1B_2)$, we get $\mu(A_1B_2) = \mu(y_2) < \mu(B_1B_2) = \mu(y_4)$, so $\mu(y_2)$ can not be equal to 1. With AGM-consistency, all happens as if an external observer deals with the possible beliefs of every player, upholding the planned equilibrium actions (as in the SE and in the ICSE) but possibly changing his view on an earlier out-of-equilibrium way of playing each time he faces a new deviation. So, if by observing that Player 1 does not play C_1, he becomes convinced that he plays B_1 ($\mu(x_3) = 1$), then, given that Player 2 plays B_2 at equilibrium, he necessarily assigns belief 1 to y_4.

We now consider Bonnano's PBE concept. The above ICSE, with $\mu(x_2) = 0.1$, $\mu(x_3) = 0.9$, $\mu(y_1) = 0$, $\mu(y_2) = 0.3$, $\mu(y_3) = 0$, and $\mu(y_4) = 0.7$ is not a PBE (Bonnano's version) in that, via Bayes' Rule, $\mu(x_2) = 0.1$ and $\mu(x_3) = 0.9$ lead to $\mu(y_1) = 0$, $\mu(y_2) = 0.1$, $\mu(y_3) = 0$ and $\mu(y_4) = 0.9$.

We finally show that, conversely, many AGM-consistent stories, and even Bonnano's PBE consistent stories, are not compatible with the concept of ICSE. So consider the game in Figure 5.2. Assume that the planned actions are the bold lines (in red) and that the beliefs (in blue) are given by $\mu(x_2) \geq 0.7$, $\mu(x_3) = 1 - \mu(x_2)$, $\mu(y_1) = \mu(x_2)$, $\mu(y_2) = 1 - \mu(x_2)$, $\mu(y_3) = \mu(y_4) = 0.5$.

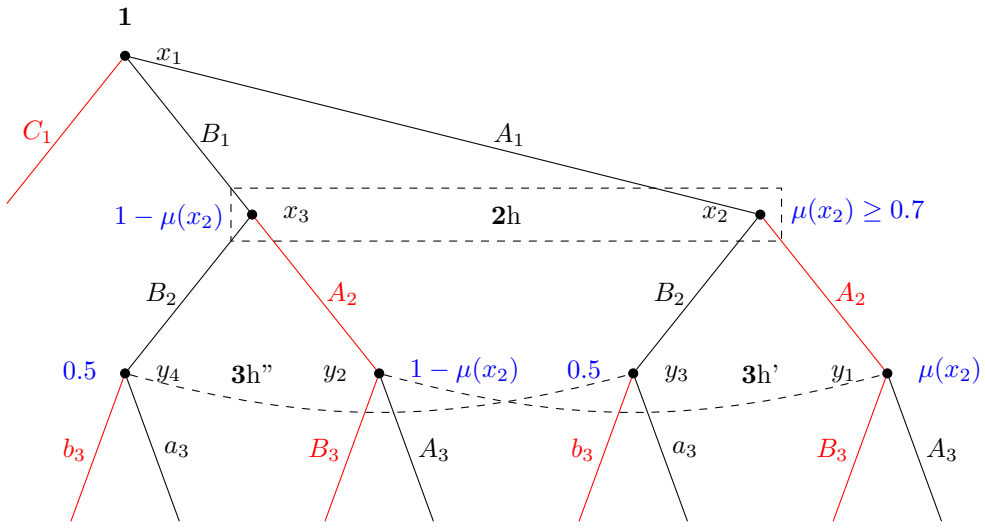


Figure 5.2 – An example illustrating the difference between ICSE and AGM-consistency.

These beliefs are AGM-consistent and they check Bonanno's PBE consistency. This is due to the fact that Bayes' Rule applies when switching from h to h' but it does not apply when switching from h to h'', since B_2 is not a plausibility-preserving

action (by contrast to A_2). An external observer, when observing the unexpected action B_2, may completely reconsider the stories of the game. At h, he believes that Player 1 more often deviates to A_1 than to B_1, but after observing the new deviation B_2, he changes his mind and thinks that Player 1 deviates to B_1 as often as to A_1. This is not possible in an ICSE because the same player, Player 3, plays at h' and h''. Regardless of Player 3's perturbations on Player 1 and Player 2's actions, we necessarily have $\mu_3(y_1) = \mu_3(y_3)$ and $\mu_3(y_2) = \mu_3(y_4)$, because Player 2 plays A_2 with the same probability at x_2 and x_3 and B_2 with the same probability at x_2 and x_3. Therefore, AGM-consistency allows an external player to have different *evolving* beliefs at a same information set (before Player 2's deviation, the external player sets $\mu(x_2) \geq 0.7$, but after Player 2's deviation he sets $\mu(x_2) = 0.5$), whereas the ICSE does not allow evolving beliefs at a same information set. Rather, it only allows different beliefs among the players (so $\mu_3(y_1) = \mu_3(y_3) = \mu(x_2)$ because these three probabilities express the way Player 3 evaluates the deviation from Player 1 towards A_1 and B_1 (before and after Player 2's choice of action), but Player 3's way of evaluating Player 1's deviations may be different from Player 2's way of evaluating these deviations, that is to say $\mu_2(x_2)$ can be different from Player 3's beliefs on x_2).

5.4 A physical distance between beliefs

While we defend the point of view that there is no logical reason that constrains people to have the same beliefs with respect to out-of-equilibrium actions, social groups often (implicitly, if not explicitly) require their members to have rather similar beliefs (Barber & Pope 2019, Gould & Klor 2019, Slothuus & Bisgaard 2021), so the pressure to modify one's beliefs is increasing in the difference between the player's beliefs and the beliefs of the group they belong to. Therefore, if all the players in a game belong to a same community, it makes sense for them to seek to reduce the distance between beliefs.

To approach the distance between beliefs, we start with a first observation. In a game, very often, the equilibrium payoffs are sustained by sets of beliefs. For example, the ICSE equilibrium payoffs (5.99, 10, 10) in the game in Figure 5.1 are sustained by Player 2's beliefs $\mu_2(x_2) \leq \frac{1}{5}$ and Player 3's beliefs $\mu_3(y_2) \geq \frac{1}{4}$. So Player 2 has to assign a probability lower than $\frac{1}{5}$ to Player 1's deviating action A_1 whereas Player 3 has to assign a probability larger than $\frac{1}{4}$ to this deviation, a fact we reproduce in Figure 5.3, in which we highlight Player 2 and Player 3's sustaining beliefs (SB) on the action A_1.

We are concerned with the closest possible beliefs sustaining an ICSE, here $\frac{1}{5}$

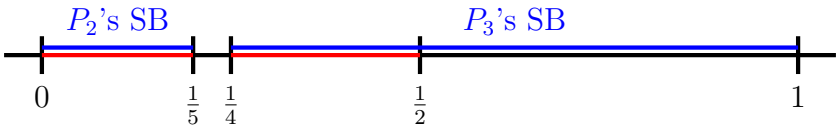


Figure 5.3 – Player 2 and Player 3's sustaining beliefs.

and $\frac{1}{4}$. Figure 5.3 illustrates three facts:

1. First, if Player 2's and Player 3's sustaining beliefs have a non-empty intersection, then common beliefs can sustain the ICSE payoffs and there exists a sequential equilibrium with these payoffs. So the minimum distance would reduce to 0.

2. Second, Player 2's and Player 3's sustaining beliefs have an empty intersection, but both sets are close ($\frac{1}{4}$-$\frac{1}{5}$ is small in comparison to 1), so that few changes in the game may lead both sets to have a non-empty intersection.

3. Third, we observe that both players can assign a probability lower than $\frac{1}{2}$ to A_1 to sustain the ICSE outcome (red probabilities for Player 2 and Player 3).

5.4.1 Ordered ICSE

Our first way to consider distance starts with the third observation. Imagine that Player 2 and Player 3 meet and discuss together: they can easily agree on the fact that both think that Player 1 deviates more often to B_1 than to A_1 (red probabilities lower than $\frac{1}{2}$). Player 2 is sure that Player 1 deviates at least 4 times more often to B_1 than to A_1. Player 3 can agree that Player 1 deviates more often to B_1 than to A_1 but at most 3 times more often. So there is a possible consensus between Player 2 and Player 3 despite the fact that they can not have the same beliefs. This consensus is on the way they order the deviations: both players believe that Player 1 more often deviates to B_1 than to A_1, yet only the "intensity" of this deviation is not shared among them.

It derives from this observation that a soft notion of proximity between beliefs consists in requiring that all the players order the perturbed strategies at each information set in the same way.

Definition 1. *An ordered ICSE is an ICSE that checks the additional condition:*

v) $\forall h \in H, \forall a, a' \in A_h, \forall j, j' \in N, \hat{\pi}^k_{j,i(h)}(a) \geq \hat{\pi}^k_{j,i(h)}(a') \Rightarrow \hat{\pi}^k_{j',i(h)}(a) \geq \hat{\pi}^k_{j',i(h)}(a')$.

In the game in Figure 5.1, it is easy to find an ordered ICSE that sustains the equilibrium actions (C_1, B_2, B_3) and that checks Definition 1. As a matter of

example, with Player 2's distribution $(\epsilon, 4\epsilon, 1-5\epsilon)$ on the actions A_1, B_1 and C_1 and with Players 3's distribution $(\epsilon, 3\epsilon, 1-4\epsilon)$ on the actions A_1, B_1 and C_1, beliefs converge to sustaining beliefs and $\epsilon = \hat{\pi}^k_{2,1(x_1)}(A_1) < 4\epsilon = \hat{\pi}^k_{2,1(x_1)}(B_1) < 1-5\epsilon = \hat{\pi}^k_{2,1(x_1)}(C_1)$ and $\epsilon = \hat{\pi}^k_{3,1(x_1)}(A_1) < 3\epsilon = \hat{\pi}^k_{3,1(x_1)}(B_1) < 1-4\epsilon = \hat{\pi}^k_{3,1(x_1)}(C_1)$, for ϵ close to 0_+.

Yet, not every ICSE equilibrium actions can be sustained by probabilities that check Definition 1. For example, if Player 3's payoff 4 after $A_1B_2B_3$ is replaced by the payoff 1.9, then (C_1, B_2, B_3) will still be an ICSE, but no ICSE checks the condition in Definition 1. This is due to the fact that we necessarily have $\hat{\pi}^k_{2,1(x_1)}(A_1) < \hat{\pi}^k_{2,1(x_1)}(B_1)$ since we need $\mu_2(x_2) \leq \frac{1}{5} < \mu_2(x_3)$, and $\hat{\pi}^k_{3,1(x_1)}(A_1) > \hat{\pi}^k_{3,1(x_1)}(B_1)$ since we need $\mu_3(y_2) \geq \frac{1}{1.9} > \frac{1}{2} > \mu_3(y_4)$.

5.4.2 How to make ICSE beliefs SE-consistent?

Our second way to consider distance consists in exploiting the small size of the interval between Player 2's sustaining beliefs and Player 3's sustaining beliefs (second observation). Clearly, with respect to the game in Figure 5.1, the ICSE payoffs could become SE payoffs (and so the minimum distance between beliefs could collapse) by changing the game in a very smooth way. By replacing the payoff 1 after $B_1B_2A_3$ by 0.87 and the payoff 6 after $B_1B_2B_3$ by 6.17, it is possible to build an ICSE that is also a SE. We get $\mu_2(x_2) = \frac{4.5}{20}$, $\mu_2(x_3) = \frac{15.5}{20}$, and $\mu_3(y_1) = 0$, $\mu_3(y_2) = \frac{4.5}{20}$, $\mu_3(y_3) = 0$, $\mu_3(y_4) = \frac{15.5}{20}$, that is to say Player 2 and Player 3 have the same beliefs on Player 1's deviations.

It follows from this observation that another way to study the proximity between beliefs consists in looking at how much we need to shake the payoffs in order to get an ICSE that is also a SE; that is, how much we should shake payoffs to get beliefs that are consistent in a SE way. In other terms, after observing that it is not possible to get a SE with the equilibrium actions of a given ICSE, we can look if small changes in payoffs can allow us to get a SE with the ICSE payoffs.

Definition 2. *The ISCE beliefs are close if they can become SE-compatible with very small changes in payoffs. In that sense, the distance in beliefs becomes the distance in payoffs required to change the ICSE equilibrium payoffs into SE equilibrium payoffs.*

The steps are the following ones. We start with ICSE equilibrium behavioral strategies that can not be part of a SE. Then we introduce variables that express changes in payoffs and we minimize the changes in payoffs under the constraint that the ICSE actions and the associated beliefs become a SE.

Let us illustrate the procedure for the game in Figure 5.1, which is represented again in Figure 5.4. A first observation is that it is always possible to change the payoffs in order to get a SE with the ICSE played actions. The optimization program makes sense only if it leads to small payoff changes. In that case, we can say that the ICSE beliefs are not much distant one from another. If so, players will not feel under pressure to change them, namely because in real life there is always some incomplete information on the exact payoffs, so the smoothly changed payoffs (needed to share the same beliefs) belong to the set of possible payoffs.

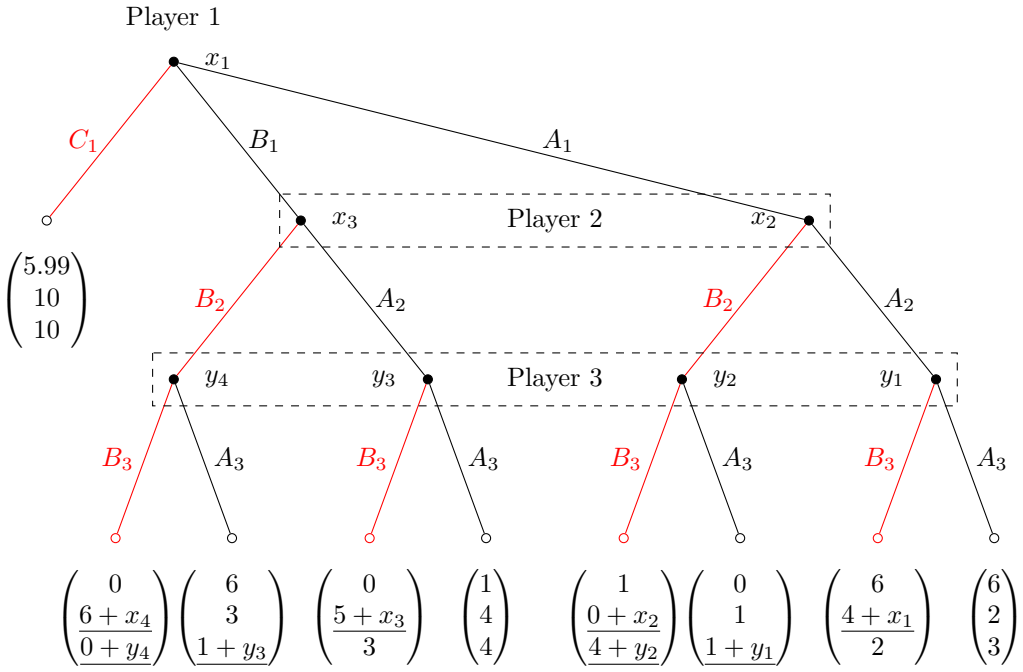


Figure 5.4 – Absolute changes in payoffs that can make ICSE beliefs SE-compatible.

A second observation is that there are only a limited number of changes to introduce in that many payoffs have no role to play in the studied equilibrium. In our game, the payoffs to work on are the ones underlined. What matters is that, for (C_1, B_2, B_3) to become a SE, the SE consistency requires that $\mu(x_2) = \mu(y_2) = \mu$ and $\mu(x_3) = \mu(y_4) = 1 - \mu$. So the program we solve is Program 1:

$$\min_{x_1,x_2,x_3,x_4,y_1,y_2,y_3,y_4,\mu} \quad {x_1}^2 + {x_2}^2 + {x_3}^2 + {x_4}^2 + {y_1}^2 + {y_2}^2 + {y_3}^2 + {y_4}^2$$

$$\text{s.t.} \quad (0 + x_2)\mu + (6 + x_4)(1 - \mu) \geq (4 + x_1)\mu + (5 + x_3)(1 - \mu) \quad \textbf{(1)}$$

$$(4 + y_2)\mu + (0 + y_4)(1_\mu) \geq (1 + y_1)\mu + (1 + y_3)(1 - \mu) \quad \textbf{(2)}$$

$$\mu \geq 0 \quad \textbf{(3)}$$

$$1 - \mu \geq 0 \quad \textbf{(4)}$$

This objective function is one among the many possible ways to measure the changes in payoffs, surely the easiest one. We will propose later a more proportional one. Equations (1) and (2) are the necessary equations ensuring sequential rationality and SE consistency. Equation (1) ensures sequential rationality for Player 2, Equation (2) ensures sequential rationality for Player 3 and sequential rationality for Player 1 is ensured in that nothing has changed for himself with respect to Figure 5.1. What matters is that conditions (1) and (2) also ensure the SE consistency, which requires that Player 2 and Player 3 put the same belief μ on x_2 (and therefore on y_2), the same belief $1-\mu$ on x_3 (and therefore on y_4) and a null belief on y_1 and y_3.

Given that the objective function goes to $+\infty$ when $\|x\|$ and/or $\|y\|$ goes to $+\infty$, and given that the admissible set is closed and that μ is limited by 0 and 1, it is easy to adapt Weierstrass' corollary to ensure that Program 1 has a global minimum. The only solution (see Appendix 3) is:

$$\begin{gathered} x_2 = -x_1 \simeq 0.0158 \\ x_4 = -x_3 \simeq 0.0564 \\ y_2 = -y_1 \simeq 0.0206 \\ y_4 = -y_3 \simeq 0.0736 \\ \mu \simeq 0.219 \\ {x_1}^2 + {x_2}^2 + {x_3}^2 + {x_4}^2 + {y_1}^2 + {y_2}^2 + {y_3}^2 + {y_4}^2 = 0.0185 \end{gathered}$$

The necessary errors are quite small, since the largest one does not exceed 0.074, which is quite small given that we work with integers ranging from 0 to 6. In other terms, we can say that our ICSE payoffs are easily SE-compatible (because the needed payoffs changes are very small). Observe that the SE belief $\mu(x_2)$ becomes 0.219, which is between $\frac{1}{5}$ and $\frac{1}{4}$.

If we switch to a more proportional way to see payoff adjustments, we can choose to switch to Figure 5.5 and to the maximization Program 2:

$$\begin{aligned} \min_{x_1,x_2,x_3,y_1,y_2,y_3,\mu} \quad & (\frac{x_1}{4})^2 + (\frac{x_2}{5})^2 + (\frac{x_3}{6})^2 + {y_1}^2 + (\frac{y_2}{4})^2 + {y_3}^2 \\ \text{s.t.} \quad & (6+x_3)(1-\mu) \geq (4+x_1)\mu + (5+x_2)(1-\mu) \quad \textbf{(1)} \\ & (4+y_2)\mu \geq (1+y_1)\mu + (1+y_3)(1-\mu) \quad \textbf{(2)} \\ & \mu \geq 0 \quad \textbf{(3)} \\ & 1-\mu \geq 0 \quad \textbf{(4)} \end{aligned}$$

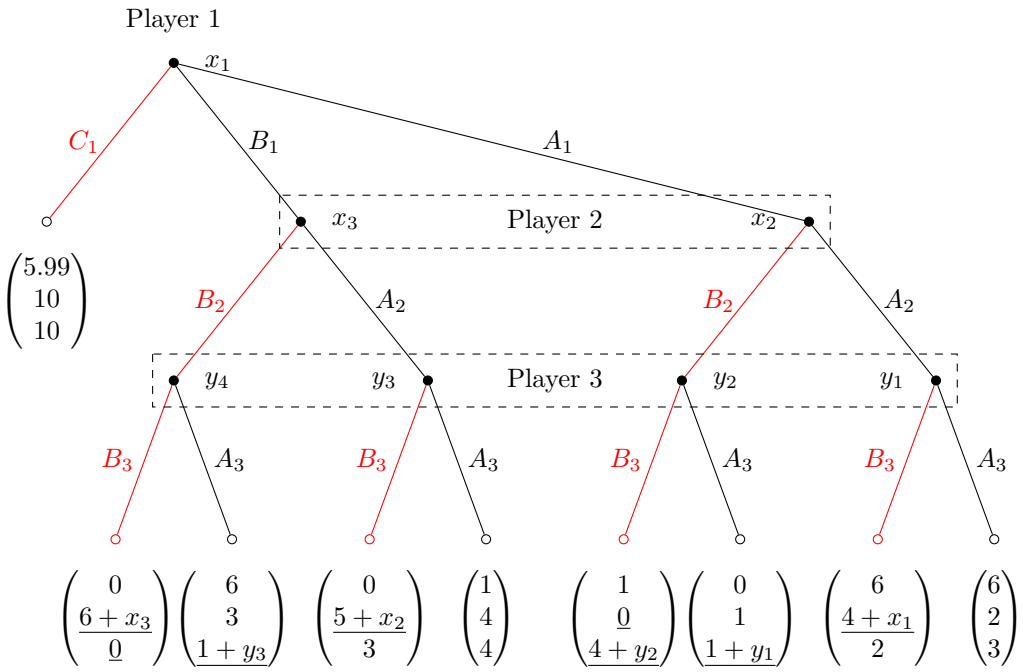


Figure 5.5 – Proportional changes in payoffs that can make ICSE beliefs SE-compatible.

The only solution (see Appendix 3) is: $x_1 = -0.0258$, $x_2 = -0.1232$, $x_3 = 0.1774$, $y_1 = -0.0021$, $y_2 = 0.0338$, $y_3 = -0.0065$ and $\mu = 0.2466$.

We can observe that $(\frac{x_1}{4})^2 + (\frac{x_2}{5})^2 + (\frac{x_3}{6})^2 + {y_1}^2 + (\frac{y_2}{4})^2 + {y_3}^2 = 0.0016$, which is again quite small, and that no term (perturbation/payoff) is larger than 0.00645. The SE belief $\mu(x_1)$ now becomes 0.247, which is again between $\frac{1}{5}$ and $\frac{1}{4}$. Several remarks have to be made.

This concept of distance is rather simple to employ given that the program is easy to write (sequential rationality and the SE consistency give the set of constraints and the objective is a function increasing in the introduced payoff perturbations). But the interpretation of the result is necessarily a little subjective. For example, in Program 2, what should we require in order to say that beliefs are close? What is the threshold $|\frac{dx}{x}|$ we should accept (where dx is the variation of payoff and x the payoff)? We can of course impose the constraint $|\frac{dx}{x}| \leq 0.1$ in order to prevent too strong payoff changes, but this gives us no way to appreciate the optimal value of the objective function. Also, should we introduce a fixed threshold on the objective function and/or the ratios $|\frac{dx}{x}|$? Or should the thresholds depend on the payoffs that divergent beliefs allow to get at equilibrium? We think that, when opting for a proportional approach (Program 2), rather than asking for $|\frac{dx}{x}| \leq 0.1$ or another small value, we should ask for a threshold whose value rises with the benefit linked to the ICSE. This way of doing is motivated by the following fact: when a player earns a large payoff, he is less induced to change things (e.g., his beliefs) and he is

less induced to ask that other persons change their beliefs.

5.4.3 Distance in ICSE beliefs and AGM-consistent beliefs

Another remark, which brings us back to Bonanno (2013, 2016)'s concept of AGM-consistency, is that this notion of distance does not take into account the number of deviations required to reach an information set. Let us consider the game in Figure 5.6. Suppose that it is possible to build an ICSE (C_1, A_2, B_3) with beliefs checking $\mu_2(x_2) \geq 0.6$ and $\mu_3(y_1) \leq 0.3$. Of course, SE consistency requires that $\mu_2(x_2) = \mu_3(y_1)$, so that we potentially have to strongly shake the payoffs in order to make the ICSE payoffs SE-compatible. In other terms, if we measure the distance as in Program 1 or Program 2, we will surely conclude that Player 2 and Player 3's beliefs are distant from one another. But this conclusion does not take into account a strong difference between h and h': h needs one deviation to be reached (Player 1's deviation) whereas h' needs two deviations to be reached (Player 1's deviation and Player 2's deviation).

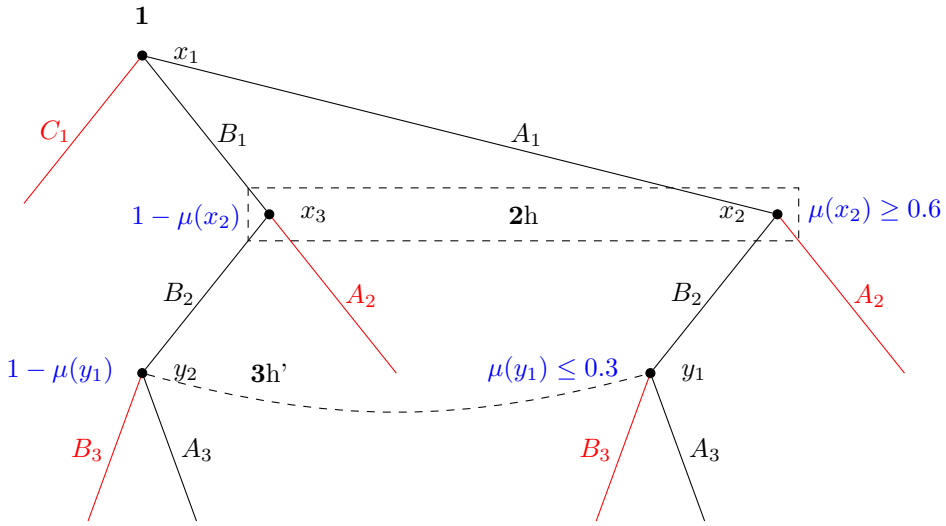


Figure 5.6 – Revisiting the links between ICSE and AGM-consistency.

This fact explains that Player 2 and Player 3's beliefs are AGM-consistent. As a matter of fact, the story A_1 is as plausible as the story A_1A_2 (because A_2 is the plausibility-preserving action), the story B_1 is as plausible as the story B_1A_2 for the same reason, and so an external observer can judge that the stories A_1 and A_1A_2 are more plausible than the stories B_1 and B_1A_2. Yet, he may also judge that the story B_1B_2 is more plausible than the story A_1A_2 because B_2 is a new deviation that totally shakes his understanding of the game: at h, after Player 1's deviation,

he thinks that Player 1 probably deviated toward A_1, so that Player 2 should play A_2, but after Player 2's unexpected deviation toward B_2 (at h'), he changes his mind and finally thinks that Player 1 probably deviated to B_1.

By contrast to Bonanno, we do not work with an external observer but only with the players themselves. This amounts, in Figure 5.6, in translating Bonanno's switch in beliefs into the following question: can we reasonably require that somebody who observes more deviations needs to share the same beliefs than somebody who observes less deviations? Player 2 observes only one deviation (Player 1's deviation) whereas Player 3 observes two deviations (Player 1's and Player 2's deviations). Facing more deviations may introduce more doubts and therefore allow for different beliefs. To say it differently, the distance in beliefs might also depend on the distance (in the number of deviations to reach it) between an information set and the equilibrium path. This amounts to saying that players facing more deviations can be expected to have more distant beliefs. In some way, if we note Player 2's payoff changes by dx, and Player 3's payoff changes by dy, it could make sense, in the game in Figure 5.6, to weight $|\frac{dx}{x}|$ more strongly than $|\frac{dy}{y}|$ in the distance function to minimize (for example $2(\frac{dx}{x})^2$ and $(\frac{dy}{y})^2$). It would perhaps even make more sense to change the power assigned to $|\frac{dx}{x}|$ and $|\frac{dy}{y}|$ given that the probability of several deviations exponentially decreases with the number of deviations (we could work with $(\frac{dx}{x})^2$ and $(\frac{dy}{y})^4$). In this way, given that $|\frac{dx}{x}|$ and $|\frac{dy}{y}|$ are lower than one, Player 3 can afford more payoff changes without increasing too much the value of the distance function: this amounts to saying that we do not judge his beliefs very distant from the other players', even if in fact they are very different.

5.5 Revealed beliefs and strategic beliefs

Taking into account the number of deviations to reach an out-of-equilibrium information set puts into light a new problem when studying the notion of distance in beliefs. What exactly is the link between beliefs and deviations? May there be a difference between *announced* beliefs and the beliefs *revealed* through the players' behavior?

Let us again consider the game in Figure 5.6. Are Player 3's beliefs really distant from Player 2's revealed beliefs? A_2 is optimal when $\mu(x_2) \geq 0.6$. Yet, when Player 3 is called on to play, Player 2 played B_2 and not A_2. Given that Player 2 plays B_2 when his beliefs check $\mu(x_2) < 0.6$ (because the complementary beliefs lead to the play of A_2), we can say that if Player 3 is called on to play (if Player 2 played B_2), then Player 2's *revealed* beliefs contradict the beliefs announced at

the information set h. And the revealed beliefs ($\mu(x_2) < 0.6$), are compatible with Player 3's beliefs, $\mu(y_1) = \mu(x_2) \leq 0.3$. So Player 3's beliefs are in reality compatible with Player 2's revealed beliefs. By the way, this might provide a logical reason for the reversal of beliefs of Bonanno's external observer at h' in the game in Figure 5.6. Given that Player 2 does not play A_2 (the action compatible with the observer's beliefs $\mu(x_2) \geq 0.6$), Player 2 reveals to the observer that his beliefs are the reversed ones, which induces the observer to change his beliefs.[5]

This way of coping with beliefs may seem attractive but it clearly leads to difficulties. First, it is often incompatible with the ICSE beliefs. In the game in Figure 5.2 for example, Player 3, with the ICSE concept, necessarily has the same beliefs at h' and h'' ($\mu(y_1) = \mu(y_3)$), so his beliefs are not reversed at h'' despite the fact that he knows, at h'', that Player 2 did not behave in conformity with his beliefs at h (given that he did not play A_2). In fact Player 3, at h'', sees Player 2's action B_2 as a trembling hand action that has no informative content and his beliefs only follow from his own perturbations on Player 1's actions A_1 and B_1. Secondly, AGM-consistent beliefs in this game are compatible with the notion of revealed beliefs, given that Player 3's beliefs at h'' take into account that Player 2, by playing B_2, revealed that his beliefs are such that $\mu(x_2) \leq 0.7$. But AGM-consistent beliefs could also assign probability 0.7 to y_3 in that the external observer is in no way compelled to change his view on Player 1's played actions after an unexpected action from Player 2.

Thirdly, taking into account revealed beliefs puts into question the measure of the distance we proposed previously. If a player tries to be close to a previous player's beliefs, then the notion of revealed beliefs requires that his beliefs must be different depending on whether the previous player played the planned action or not. In Figure 5.2 for example, a SE requires $\mu(x_2) = \mu(y_1) = \mu(y_3)$, so if an equilibrium starts with $\mu(y_1) \neq \mu(y_3)$, the distance is necessarily strictly positive despite the fact that revealed beliefs by definition require $\mu(y_1) \neq \mu(y_3)$. This suggests that taking into account the number of deviations required to reach an information set must change our measure of distance.

Finally, we should consider the notion of revealed beliefs with suspicion. Let us consider the game in Figure 5.7.

In every sequential equilibrium, we have $\mu(x_2) = \mu(y_1)$, and it follows that Player 2 always plays B_2 (because either $\mu(x_2) = \mu(y_1) > \frac{1}{2}$, Player 3 plays A_3 and therefore Player 2 plays B_2 or $\mu(x_2) = \mu(y_1) < \frac{1}{2}$, Player 3 plays B_3 and therefore

[5]But Bonanno (2013, 2016) does not require this reversal. AGM-consistent assessments also allow beliefs such as $\mu(x_2) = \mu(y_1)$ and $\mu(x_3) = \mu(y_2)$.

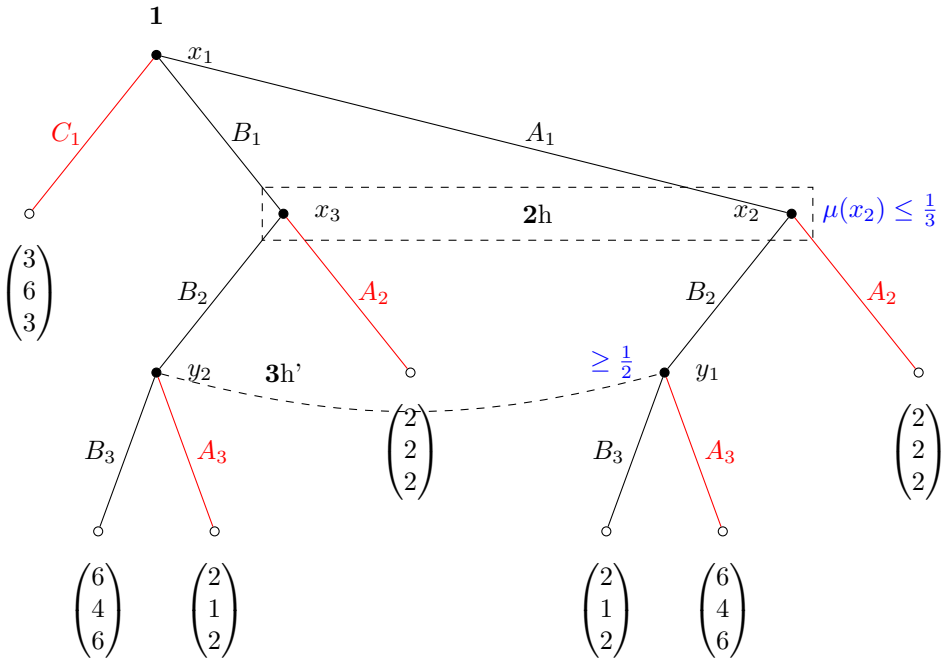


Figure 5.7 – Revealed and announced beliefs.

Player 2 plays B_2, or $\mu(x_2) = \mu(y_1) = \frac{1}{2}$, Player 3 plays A_3 with probability q and therefore Player 2 plays B_2). This implies that Player 1 plays A_1 if $\mu(x_2) = \mu(y_1) > \frac{1}{2}$, B_1 if $\mu(x_2) = \mu(y_1) < \frac{1}{2}$, and A_1 or B_1 depending on the value of q if $\mu(x_2) = \mu(y_1) = \frac{1}{2}$. Hence, both Players 1 and 3 get 6 at any sequential equilibrium and Player 2 gets at best 4.[6]

Now consider the ICSE given in Figure 5.7, with the SE incompatible beliefs $\mu(x_2) \leq \frac{1}{3}$ and $\mu(y_1) \geq \frac{1}{2}$. This new profile of actions and beliefs checks AGM-consistency (and Bonanno's PBE consistency) and is conform to revealed beliefs. As a matter of fact, given his beliefs and Player 3's action A_3, Player 2 should play A_2, which he does at equilibrium and which induces Player 1 to play C_1. So, if Player 3 is called on to play, this means that Player 2 played B_2 which, according to revealed beliefs, reveals that he does not believe that Player 1 played A_1 with a probability lower than $\frac{1}{3}$. Therefore, Player 3, if he wishes to share similar beliefs to Player 2, can believe that Player 1 played A_1 with a probability larger than $\frac{1}{3}$ and possibly larger than $\frac{1}{2}$. These beliefs induce him to play A_3. So revealed beliefs sustain a profile of strategies where Players 1 and 3 only get 3 and Player 2 gets 6, by leading Player 1 to play C_1.

But let us look more closely at Player 2's revealed beliefs. By playing B_2 instead of the planned action A_2, Player 2 sends the following message to Player 3:

[6]There also exists a SE where Player 1 plays A_1 and B_1 with probability $\frac{1}{2}$ and Player 3 plays A_3 and B_3 with probability $\frac{1}{2}$, so they both get 4 given that Player 2 plays B_2.

"Normally I play A_2 because it is my best response to your action A_3, since I believe that Player 1 much more deviates to B_1 than to A_1; so, if you see me playing B_2, this means that I changed my opinion about Player 1's deviation, so that you are right in believing that he more often deviated to A_1, and so you are right in playing A_3." The problem is that this message is at best "cheap talk" because Player 2 has no information to reveal to Player 3. Player 2 has no idea about the potential deviations of Player 1, given that Player 1 does not deviate when he expects Player 2 to play A_2. One more time, by contrast to Aumann (1976), Geanakoplos & Polemarchakis (1982), and Hart & Tauman (2004), the out-of-equilibrium actions played by the players do not provide information on a true state to discover but on an action that will never be observed at equilibrium. Player 1 does not deviate and so there is nothing to learn about his deviation. If Player 3 "naively" believes that Player 2 can reveal something about Player 1's deviations with his behavior, then he gives Player 2 the power of manipulating Player 3's beliefs to his advantage. As a matter of fact, the beliefs in this ICSE clearly are advantaging Player 2 because they induce Player 1 to play C_1, so they lead to the payoff profile (3,6,3) which is exclusively in advantage of Player 2.

5.6 Discussion

Our analysis suggests that out-of-equilibrium beliefs might be here to justify the players' behaviour at out-of-equilibrium information sets. Choosing them in a given way may help to get a higher payoff, which means that they belong to the strategy set. Let us recall that in a SE, everybody builds the beliefs on a same profile of perturbations, so a player does not really choose his beliefs given that he applies Bayes' Rule to the same perturbations. So in a SE, out-of-equilibrium beliefs do not belong to the player's strategy set. By contrast, in an ICSE, each player chooses his own profile of perturbations, which means that he chooses his own beliefs at out-of-equilibrium sets. This degree of liberty can be exploited to build beliefs that lead to interesting payoffs. In the game in Figure 5.7, it is good for Player 2 to build beliefs (about Player 1's deviation) that are strikingly different from those of Player 3 in order to justify the action A_2 that prevents Player 1 from deviating from C_1, the most interesting action for Player 2. By contrast, for Player 3, it is better to have similar beliefs than Player 2, in order to lead Player 2 to play B_2, which ultimately leads to the payoffs (6,4,6). So we are tempted to say that, given that each player chooses his set of perturbations, out-of-equilibrium beliefs belong to the strategy set. This fact induces two consequences: we have to reconsider the notion

of sequential rationality and we have to reconsider the notion of distance between beliefs.

We start with the notion of distance by coming back to the game studied in Figure 5.1. We already made the observation, in Section 3, when opting for a proportional approach (Program 2), that rather than asking for $|\frac{dx}{x}| \leq 0.1$ or another threshold, we should ask for a threshold whose value rises with the benefit linked to the ICSE payoffs. As a matter of fact, if everybody benefits from the ICSE payoffs then nobody cares about the distance between the beliefs necessary to sustain the equilibrium. By contrast, in the game in Figure 5.7, Player 3 may require that Player 2 has beliefs that are close to his own beliefs (especially if Player 2 is a newcomer in the community) because similar beliefs among Player 2 and Player 3 are necessary for Player 3 to get the nice payoff 6. Conversely, if Player 3 is the newcomer in the community, then Player 2 might not pressure Player 3 to adopt beliefs close to his own. The (social) pressure to modify beliefs so as to be close to another player's beliefs must therefore be contrasted with the benefits (in terms of actions played) of holding different beliefs.

Concerning sequential rationality, given that each player chooses his perturbation scheme, these perturbation schemes belong to his strategy set, that is, each player will build (in a consistent way) beliefs at his out-of-equilibrium information sets to get a better payoff. This changes nothing with respect to the definition of individual consistency, but this should lead us to reconsider Kreps & Wilson (1982)'s notion of sequential rationality. In some way, we should add that, for each player, given the strategies played by the other players, there does not exist a perturbation scheme that is sequentially rational (as defined by Kreps & Wilson 1982), individually-consistent and that leads to a larger payoff for the player.

However, sequential rationality rests on unilateral deviations and this additional condition might thus not always help. For example, the ICSE in Figure 5.7 would resist such an additional condition despite the fact that Player 3 would like to adapt his beliefs to Player 2's beliefs to compel him to play B_2. The problem is that, as long as Player 2 plays A_2 and Player 1 plays C_1, Player 3's beliefs and actions have no impact on his equilibrium payoff. The idea is that each player selects the perturbation scheme associated to the ICSE that leads him to his largest equilibrium payoff. If so, in the game in Figure 5.7, Player 3 should opt for a perturbation scheme (on Player 1's actions) similar to Player 2's, to push Player 1 and Player 2 to play A_1 and B_2 for example (he can choose the SE (A_1, B_2, A_3) with the beliefs $\mu(x_2) = \mu(y_1) = 1$). Yet, Player 2 would of course choose another perturbation scheme, namely the one leading to the ICSE in Figure 5.7. So this

additional condition, in the game in Figure 5.7, leads to the non-existence of a system of perturbation schemes both selected by Player 2 and Player 3. In the game in Figure 5.1, Player 2 and Player 3 can select the same perturbation scheme, namely the one of an ICSE leading to the payoffs (5.99, 10,10). Player 1 can not counter Player 2 and Player 3's selection, in that if they play B_2 and B_3 he is constrained to play C_1. Yet, even in this game, Player 1 may opt for another equilibrium in which he plays A_1, therefore constraining Player 2 and Player 3 to play A_2 and A_3. Therefore, there is no obvious way, even if we switch to coalitions of players, to clearly formalize and express the wish to select payoff-optimizing perturbation schemes. The only trivial configuration is a game such that one ICSE ensures the best payoff to all the players, so that the grand coalition of all the players will be incentivized to select it.

5.7 Conclusion

The chapter started with an obvious observation: there is no reason that leads every player to build the perturbed strategies similarly. Each player has to respect the probabilities assigned to actions that are in the support of the equilibrium, but, given that there does not exist an external observer who can decide for the profiles of ϵ-perturbations, each player is free to build the perturbations assigned to the actions out of the support of the equilibrium. It follows that in an ICSE, players can have different beliefs at out-of-equilibrium information sets.

This led us first to evaluate the distance between different beliefs, because players in the same community are often expected to share similar beliefs. We did this in Section 3 in two ways: (i) an ordering of perturbations and (ii) the minimization of changes in payoffs necessary to make the ICSE beliefs SE-compatible.

We then focused on the function held by beliefs at out-of-equilibrium sets. Since players can build their beliefs at out-of-equilibrium sets, they might build them strategically in order to improve their payoffs. This observation led us to reconsider the traditional concept of sequential rationality, by further requiring that there does not exist a perturbation profile that is individually-consistent and that provides greater payoffs to the player, even though such an additional constraint is not always easy to cope with.

5.8 Appendix 1

We show that there are no sequential equilibria supporting the action C_1 for player 1 (and therefore the socially optimal situation) in the game shown in Figure 5.1.

Case 1: Player 2 plays A_2. Therefore player 1 plays A_1.

Case 2: Player 2 plays B_2.

(i): If player 3 plays A_3, player 1 plays B_1.

(ii): If player 3 plays B_3, then player 1 might want to play C_1, but we have already shown that the beliefs that would support this equilibrium are not mutually consistent.

(iii): If player 3 plays A_3 and B_3, then $\gamma = \frac{1}{4}$, and so necessarily $\alpha = \frac{1}{4}$. But then, player 2 would prefer to play A_2. To show this, first let r be the probability that player 3 plays A_3. By playing A_2, player 2's expected payoff is $\frac{1}{4}(2r+4(1-r)+\frac{3}{4}(4r+5(1-r)) = \frac{1}{4}(14r+19(1-r))$. By playing B_2, player 2's expected payoff is $\frac{1}{4}r+\frac{3}{4}(3r+6(1-r)) = \frac{1}{4}(10r+18(1-r))$, which is strictly inferior to the expected gain of playing A_2.

Case 3: Player 2 plays A_2 and B_2.

(i): Player 3 plays A_3. In this case, it would not be profitable for player 2 to randomize, given that playing only A_2 would allow her to always gain strictly more.

(ii): Player 3 plays B_3. Given player 2's indifference between A_2 and B_2, it is necessary that $\alpha = \frac{1}{5}$. To show that with these beliefs, player 3 would want to deviate, first note q the probability that player 2 would play A_2. Then by playing A_3, player 3's expected gain would be $\frac{1}{5}(1+2q)+\frac{4}{5}(1+3q) = \frac{1}{5}(5+14q)$. By playing B_3, player 3's expected gain would be $\frac{1}{5}(4-2q)+\frac{4}{5}(3q) = \frac{1}{5}(4+10q)$, which is strictly inferior to the expected gain player 3 would receive by playing A_3.

(iii): Player 3 plays A_3 and B_3. Let r be the probability that player 3 plays A_3, and q the probability that player 2 plays A_2. Let ϵ_0 and ϵ_1 be the perturbations associated to A_1 and B_1 respectively. The expected gain of playing A_2 for player 2 is $\epsilon_0(2r+4(1-r))+\epsilon_1(4r+5(1-r)) = \epsilon_0(4-2r)+\epsilon_1(5-r)$. The expected gain of playing B_2 for player 2 is $\epsilon_0 r+\epsilon_1(3r+6(1-r)) = \epsilon_0 r+\epsilon_1(6-3r)$. Equalizing these expected gains yields $\epsilon_0(4-2r)+\epsilon_1(5-r) = \epsilon_0 r+\epsilon_1(6-3r)$, or $\epsilon_0(4-3r) = \epsilon_1(1-2r)$.

The expected gain of playing A_3 for player 3 is $\epsilon_0(3q+(1-q))+\epsilon_1(4q+(1-q)) = \epsilon_0(1+2q)+\epsilon_1(1+3q)$. The expected gain of playing B_3 for player 3 is $\epsilon_0(2q+4(1-q))+\epsilon_1 3q = \epsilon_0(4-2q)+\epsilon_1 3q$. Equalizing these expected gains yields $\epsilon_0(1+2q)+\epsilon_1(1+3q) = \epsilon_0(4-2q)+\epsilon_1 3q$, or $\epsilon_0(3-4q) = \epsilon_1$. It follows that $\epsilon_1 = \epsilon_0(3-4q) = \epsilon_0\frac{4-3r}{1-2r}$. Therefore, $3-4q = \frac{4-3r}{1-2r}$, so $4q = 3-\frac{4-3r}{1-2r} = \frac{3-6r-4+3r}{1-r} = \frac{-1-3r}{1-r}$, which is strictly inferior to 0; an impossible event.

5.9 Appendix 2

Proof of Proposition 5.3.1. 1. This follows from the fact that in an ICSE, all strategies are sequentially rational and beliefs are obtained via Bayes' Rule applied to strategies close to the true ones (even if the perturbations are not the same among players). In an ICSE, players agree on the planned actions, even those at unreached subgames, so the ICSE induces a Nash equilibrium in each subgame.

2. This follows directly from the definition of both concepts.

3. This follows from the fact that the perturbations required by Player 2 (to build his beliefs) are about actions played by Player 1, and the perturbations required by Player 1 are about actions played by Player 2. Both players do not work with different perturbations about actions played by another (third) player. So we can work with one set of perturbations for the game, which is the same for both players (by taking Player 1's perturbations (about Player 2's actions) and Player 2's perturbations (about Player 1's actions)).

4. To show why an ICSE is not necessarily a PBE, we choose a game closer to the games studied by Fudenberg & Tirole (1991), by changing our main example in the following way. θ_1 and θ_1' are Player 1's two possible types, unknown to Player 2 and to Player 3 (prior probabilities ρ and $1-\rho$). So we get the game in Figure 5.8.

   According to Fudenberg & Tirole (1991)'s PBE equilibrium concept, Player 2 and Player 3 share the same beliefs everywhere (Condition B(iv) p.332), and these beliefs are build using the history of play whenever possible. So, if $\mu(x_4) = \mu_2(\theta_1'/h) = \mu(\theta_1'/h) = \frac{1}{5}$, we get:

$$\begin{aligned}\mu(y_3) &= \frac{\mu(\theta_1'/h)\pi_2(B_2)}{\mu(\theta_1/h)\pi_2(B_2) + \mu(\theta_1/h)\pi_2(A_2) + \mu(\theta_1'/h)\pi_2(B_2) + \mu(\theta_1'/h)\pi_2(A_2)} \\ &= \mu(\theta_1'/h) \\ &= \frac{1}{5},\end{aligned}$$

   (due to the Condition B(ii) p.332). Therefore, we can not get $\mu(y_3) = \frac{1}{4}$. Player 3's beliefs are built like Player 2's. Given that B_2 is an expected action, and given that the beliefs at h are not 0, the beliefs at y_3 are necessarily the same than the ones at x_4.

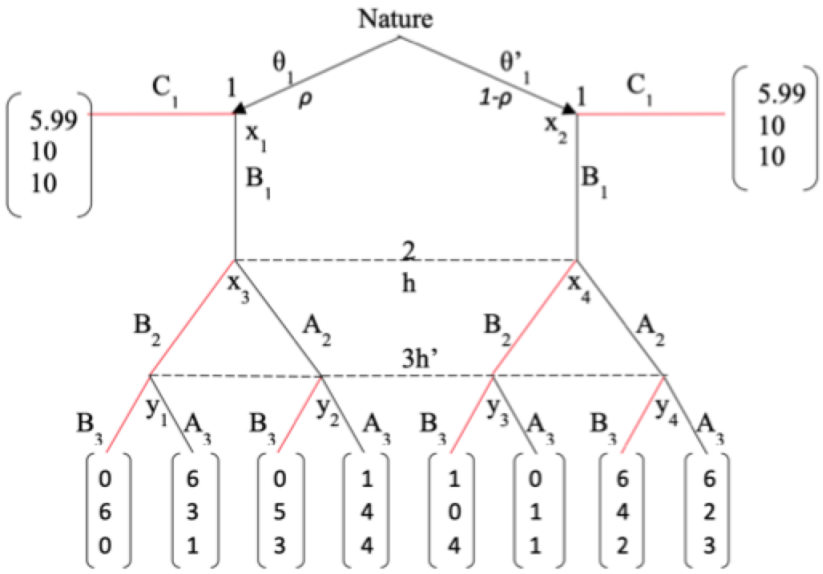


Figure 5.8 – Distinction between ICSE and PBE (1/3).

The ICSE concept takes into account that B_2 is an expected action, but it allows Player 3 not to share Player 2's beliefs at h. Therefore, we keep Condition B(ii) but not Condition B(iv), in that Player 2 (respectively Player 3) assigns probability $\frac{1}{5}$ (respectively $\frac{1}{4}$) to θ' if h is reached.

Yet, all PBE are not necessarily ICSE. For example, consider Fudenberg & Tirole (1991)'s example reproduced in Figure 5.9 and Figure 5.10 (Figure 8.9 p.346 in their book). The beliefs are in red. The beliefs assigned to the states θ have been obtained by Bayesian inference from previous play. Player 1 is the player who plays the actions a_1^*, a_1' and a_1'', while e_k, e_k' are perturbations going to 0.

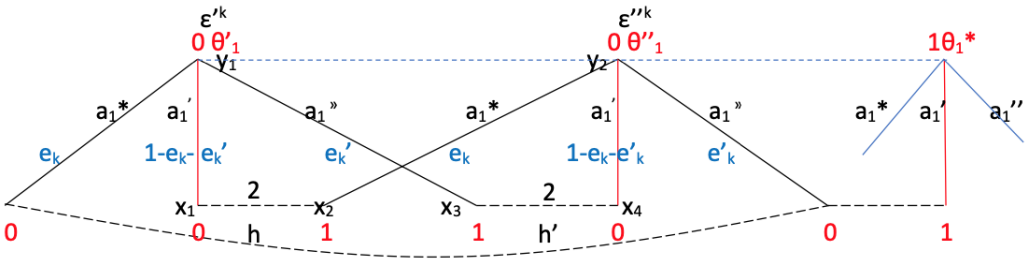


Figure 5.9 – Distinction between ICSE and PBE (2/3).

Fudenberg & Tirole (1991) say that the beliefs (in red) at h and h' belong to a PBE because the PBE places no restrictions on the beliefs at h and h',

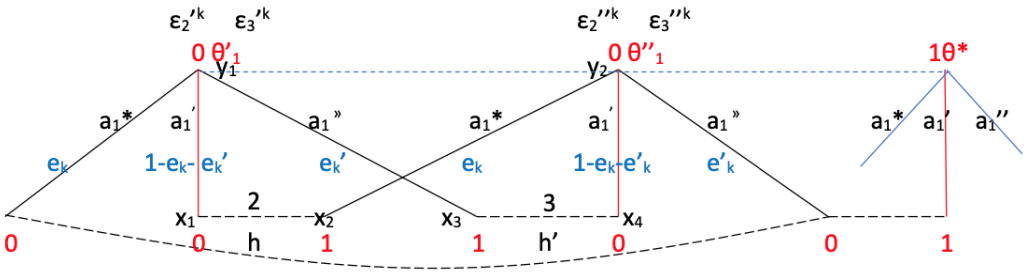


Figure 5.10 – Distinction between ICSE and PBE (3/3).

since both beliefs that lead to h and h' are 0. The only condition is that these beliefs have to be common to all players.

Yet the beliefs at h and h' can not belong to an ICSE when the same player, Player 2 in Figure 5.9, plays at h and h'. As a matter of fact, by calling $\epsilon^{'k}$ and $\epsilon^{''k}$ the probabilities to reach y_1 and y_2, respectively, which go to 0 given the beliefs, we get $\mu_2(x_3) = \lim_{\epsilon \to 0}(\frac{\epsilon^{'k} e_k'}{\epsilon^{'k} e_k' + \epsilon^{''k}(1-e_k-e_k')})$, which can go to 1 only if $\frac{\epsilon^{''k}(1-e_k-e_k')}{\epsilon^{'k} e_k'}$ goes to 0, which requires that $\frac{\epsilon^{''k}}{\epsilon^{'k}} \to 0$. But then $\mu_2(x_2) = \lim_{\epsilon \to 0}(\frac{\epsilon^{''k} e_k}{\epsilon^{''k} e_k + \epsilon^{'k}(1-e_k-e_k')}) \to 0$. So the PBE is not a ICSE.

By contrast, when there are two different players at h and h'', like in Figure 5.10 (Player 2 and Player 3 respectively), the PBE is an ICSE, since we can take different perturbations leading to y_1 and y_2 for Player 2 ($\epsilon_2^{'k}$ and $\epsilon_2^{''k}$) and Player 3 ($\epsilon_3^{'k}$ and $\epsilon_3^{''k}$), respectively. Therefore, we get: $\mu_3(x_3) = \lim_{\epsilon \to 0}(\frac{\epsilon_3^{'k} e_k'}{\epsilon_3^{'k} e_k' + \epsilon_3^{''k}(1-e_k-e_k')})$, which can go to 1 if $\frac{\epsilon_3^{''k}}{\epsilon_3^{'k} e_k'} \to 0$. And we get $\mu_2(x_2) = \lim_{\epsilon \to 0}(\frac{\epsilon_2^{''k} e_k}{\epsilon_2^{''k} e_k + \epsilon_2^{'k}(1-e_k-e_k')})$, which can also go to 1 if $\frac{\epsilon_2^{'k}}{\epsilon_2^{''k} e_k} \to 0$. Given that the set of perturbations is different for the two players, both conditions can be fulfilled and the PBE becomes an ICSE.

5. To formally describe the concept of SCE, we need to introduce some more notation. Let a mixed strategy profile be represented by $\sigma = (\sigma_i)_{i \in N}$, while a behavioral strategy profile will be denoted, as before, by $\pi = (\pi_i)_{i \in N}$. We refer to information sets that are reached with positive probability under the mixed strategy profile σ as $\bar{H}(\sigma)$, and write h_i for a specific information set controlled by player i. We write the behavioral representation of a mixed strategy σ_i as $\hat{\pi}_i(\cdot|\sigma_i)$, such that $\hat{\pi}_i(h_i|\sigma_i)$ represents the probability distribution over actions induced by the mixed strategy σ_i at information set h_i. Furthermore, let μ_i represent player i's beliefs about their opponents' play, such that μ_i is a probability distribution over Π_{-i}, the set of other players' behavioral strategies, with typical element π_{-i}.

A (mixed) strategy profile σ is a *Self-Confirming Equilibrium* (Fudenberg & Levine 1993) if, $\forall i \in N$ and $\forall s_i \in support(\sigma_i)$, there exists beliefs μ_i such that:

(a) s_i maximizes $u_i(\cdot, \mu_i)$, and

(b) $\mu_i[\{\pi_{-i}|\pi_j(h_j) = \hat{\pi}_j(h_j|\sigma_j)\}] = 1, \forall j \neq i, \forall h_j \in \bar{H}(s_i, \sigma_{-i})$.

In words, every players' subjective probability distribution needs to put probability 1 to strategy profiles that are compatible with observed play (reached with positive probability). That is, players' expectations need to be right on the equilibrium path, but need not be right at information sets that are never reached.

Importantly, since each player has to best respond only to the *observed* actions of other players, the SCE only requires that players play a best response to their *beliefs* about other player's actions out of the equilibrium strategy path. This implies that a SCE is not necessarily an ISCE, since the ICSE requires that players play best responses to other player's *actions*, even at information sets that are out of the equilibrium strategy path. On the other hand, an ICSE is always a SCE, since it is always possible to create a SCE in which players play best responses, and have accurate beliefs, even at out-of-equilibrium information sets.

■

Proof of Proposition 5.3.2. 1. AGM-consistency works with plausibility orders on stories. It is a qualitative notion that focuses on plausibility-preserving actions, which are actions played with a positive probability in the game. Given that the ICSE also respects the actions played with a positive probability in the game, an ICSE is usually AGM-consistent, except if it leads to 0-1 assessments.

2. Bonanno's PBE concept transforms the qualitative notion into a quantitative one, by requiring that the beliefs respect Bayes' Rule when it applies, i.e., in presence of plausibility-preserving actions. Given that Bonanno's concept only introduces one probability distribution at each out-of-equilibrium information set, it immediately follows that it cannot intersect with the ICSE concept, which works with the Bayesian rule applied to several perturbation distributions, one for each player.

■

5.10 Appendix 3

Program 1:

$$\begin{aligned}
\min_{x_1,x_2,x_3,x_4,y_1,y_2,y_3,y_4,\mu} \quad & {x_1}^2 + {x_2}^2 + {x_3}^2 + {x_4}^2 + {y_1}^2 + {y_2}^2 + {y_3}^2 + {y_4}^2 \\
\text{s.t.} \quad & (0+x_2)\mu + (6+x_4)(1-\mu) \geq (4+x_1)\mu + (5+x_3)(1-\mu) \quad \textbf{(1)}\ \lambda_1 \\
& (4+y_2)\mu + (0+y_4)(1-\mu) \geq (1+y_1)\mu + (1+y_3)(1-\mu) \quad \textbf{(2)}\ \lambda_2 \\
& \mu \geq 0 \quad \textbf{(3)}\ \lambda_3 \\
& 1-\mu \geq 0 \quad \textbf{(4)}\ \lambda_4
\end{aligned}$$

Equations **(1)** and **(2)** can be rewritten:

$$\begin{aligned}
1 - x_3 + x_4 - \mu(5 + x_1 - x_2 - x_3 + x_4) \geq 0 \quad \textbf{(1)} \\
-1 - y_3 + y_4 - \mu(-4 + y_1 - y_2 - y_3 + y_4) \geq 0 \quad \textbf{(2)}
\end{aligned}$$

The KT function becomes:

$${x_1}^2 + {x_2}^2 + {x_3}^2 + {x_4}^2 + {y_1}^2 + {y_2}^2 + {y_3}^2 + {y_4}^2 - \lambda_1(1 - x_3 + x_4 - \mu(5 + x_1 - x_2 - x_3 + x_4))$$
$$- \lambda_2(-1 - y_3 + y_4 - \mu(-4 + y_1 - y_2 - y_3 + y_4)) - \lambda_3\mu - \lambda_4(1-\mu).$$

The KT equations are:

$$\begin{aligned}
2x_1 + \lambda_1\mu = 0 \quad \textbf{(a)} \\
2x_2 - \lambda_1\mu = 0 \quad \textbf{(b)} \\
2x_3 + \lambda_1(1-\mu) = 0 \quad \textbf{(c)} \\
2x_4 - \lambda_1(1-\mu) = 0 \quad \textbf{(d)} \\
2y_1 + \lambda_2\mu = 0 \quad \textbf{(e)} \\
2y_2 - \lambda_2\mu = 0 \quad \textbf{(f)} \\
2y_3 + \lambda_2(1-\mu) = 0 \quad \textbf{(g)} \\
2y_4 - \lambda_2(1-\mu) = 0 \quad \textbf{(h)} \\
\lambda_1(5 + x_1 - x_2 - x_3 + x_4) + \lambda_2(-4 + y_1 - y_2 - y_3 + y_4) - \lambda_3 + \lambda_4 = 0 \quad \textbf{(i)}
\end{aligned}$$

It follows that $x_2 = -x_1 \geq 0$, $x_4 = -x_3 \geq 0$, $y_2 = -y_1 \geq 0$, and $y_4 = -y_3 \geq 0$, due to the positivity of the KT multipliers.

Both Conditions **(1)** and **(2)** are necessarily checked with equality, because $\lambda_1 = 0$ leads to $\lambda_2(-4 - 2y_2 + 2y_4) = 0$, hence $y_4 \geq 2$, which can clearly not lead to a global minimum given our numerical introduction, and $\lambda_2 = 0$ leads to

$\lambda_1(5-2x_2+2x_4)=0$, hence $x_4 \geq 2.5$, which can not lead to a global minimum for the same reason.

We seek a solution that checks $0<\mu<1$, and so $\lambda_3=\lambda_4=0$. It follows that $\lambda_1=\frac{2x_2}{\mu}=\frac{2x_4}{1-\mu}$, hence $\mu=\frac{x_2}{x_2+x_4}$, and $\lambda_2=\frac{2y_2}{\mu}=\frac{2y_4}{1-\mu}$, hence $\mu=\frac{y_2}{y_2+y_4}$. As a result:

$$\begin{aligned} x_2y_4 &= y_2x_4 \quad \textbf{(j)} \\ 2(x_2^2+x_4^2) &= 4x_2-x_4 \quad \textbf{(k)} \\ 2(y_2^2+y_4^2) &= -3y_2+y_4 \quad \textbf{(l)} \\ x_2(1+2x_4)+y_2(-1+2y_4) &= 0 \quad \textbf{(m)} \end{aligned}$$

The only solution is:

$$\begin{aligned} x_2=-x_1 &= -\frac{53\sqrt{29}-290}{290} \simeq 0.0158 \\ x_4=-x_3 &= \frac{33\sqrt{29}-145}{580} \simeq 0.0564 \\ y_2=-y_1 &= \frac{83\sqrt{29}-435}{580} \simeq 0.0206 \\ y_4=-y_3 &= -\frac{19\sqrt{29}-145}{580} \simeq 0.0736 \\ \mu &= \frac{-106\sqrt{29}+580}{435-73\sqrt{29}} \simeq 0.219 \\ \lambda_1 &= \frac{435-73\sqrt{29}}{290} \simeq 0.1444 \\ \lambda_2 &= \frac{64\sqrt{29}-290}{290} \simeq 0.1885. \end{aligned}$$

Program 2:

$$\begin{aligned} \min_{x_1,x_2,x_3,y_1,y_2,y_3,\mu} \quad & (\frac{x_1}{4})^2+(\frac{x_2}{5})^2+(\frac{x_3}{6})^2+{y_1}^2+(\frac{y_2}{4})^2+{y_3}^2 \\ \text{s.t.} \quad & (6+x_3)(1-\mu) \geq (4+x_1)\mu+(5+x_2)(1-\mu) \quad \textbf{(1)}\ \lambda_1 \\ & (4+y_2)\mu \geq (1+y_1)\mu+(1+y_3)(1-\mu) \quad \textbf{(2)}\ \lambda_2 \\ & \mu \geq 0 \quad \textbf{(3)}\ \lambda_3 \\ & 1-\mu \geq 0 \quad \textbf{(4)}\ \lambda_4 \end{aligned}$$

Equations **(1)** and **(2)** can be rewritten:

$$\begin{aligned} 1+x_3-x_2+\mu(-5-x_1+x_2-x_3) &\geq 0 \quad \textbf{(1)} \\ -1-y_3+\mu(4-y_1+y_2+y_3) &\geq 0 \quad \textbf{(2)} \end{aligned}$$

The KT function becomes:

$$(\frac{x_1}{4})^2 + (\frac{x_2}{5})^2 + (\frac{x_3}{6})^2 + {y_1}^2 + (\frac{y_2}{4})^2 + {y_3}^2 - \lambda_1(1 + x_3 - x_2 + \mu(-5 - x_1 + x_2 - x_3))$$
$$- \lambda_2(-1 - y_3 + \mu(4 - y_1 + y_2 + y_3)) - \lambda_3\mu - \lambda_4(1 - \mu).$$

The KT equations are:

$$\frac{2x_1}{16} + \lambda_1\mu = 0 \quad \textbf{(a)}$$
$$\frac{2x_2}{25} + \lambda_1(1 - \mu) = 0 \quad \textbf{(b)}$$
$$\frac{2x_3}{36} - \lambda_1(1 - \mu) = 0 \quad \textbf{(c)}$$
$$2y_1 + \lambda_2\mu = 0 \quad \textbf{(e)}$$
$$\frac{2y_2}{16} - \lambda_2\mu = 0 \quad \textbf{(f)}$$
$$2y_3 + \lambda_2(1 - \mu) = 0 \quad \textbf{(g)}$$
$$\lambda_1(5 + x_1 - x_2 + x_3) + \lambda_2(-4 + y_1 - y_2 - y_3) - \lambda_3 + \lambda_4 = 0 \quad \textbf{(i)}$$

We look for a solution such that Conditions **(1)** and **(2)** are checked with equality, and such that $0 < \mu < 1$ (so that $\lambda_3 = \lambda_4 = 0$. The only solution gives: $x_1 = -0.0258$, $x_2 = -0.1232$, $x_3 = 0.1774$, $y_1 = -0.0021$, $y_2 = 0.0338$, $y_3 = -0.00645$, $\lambda_1 = 0.01308$, $\lambda_2 = 0.01712$ and $\mu = 0.24657$.

General conclusion

The principal objective of this thesis has been to reevaluate the "behavioral" turn in economics and game theory. This "behavioral" turn, inspired by research in psychology, was motivated by the observation that individuals often do not behave in accordance with the predictions of standard economics and game-theoretic models based on self-interest maximization and Bayesian rationality. Rather, individuals often adopt beliefs that are not justified by the evidence, take the welfare of others into account in social dilemmas, and more generally seem to rely more on emotions than "rationality" when making decisions. In this thesis, I proposed that the deviations from the predictions of standard models can often be explained using standard economic and game-theoretic tools, provided that researchers take into account the social incentives that individuals face. The argument I tried to advance is that our psychological and emotional mechanisms are not exogenously biasing our decisions but that they themselves respond to incentives that may not be readily apparent. Therefore, tools based on individual rationality and self-interest maximization can be fruitfully applied in understanding how these mechanisms respond to incentives.

In Chapter 1, I argued that the *feeling* of wanting to improve one's self-image and the subsequent *need* to self-signal themselves require a more ultimate explanation, since both—undertaken for their own sake—are purely inconsequential. Given that learning and emotional mechanisms track rewards and punishments in our environment, the positive feeling associated with a positive self-image must necessarily stem from material or social benefits associated with having a positive image. Alternatively, the negative feeling associated with behaving inconsistently or undertaking a socially disapproved behavior must necessarily stem from material or social costs associated with these behaviors. This led me to conclude that the desired self-image must represent the desired social image, in the sense that individuals are ultimately motivated to improve and maintain their reputation in the eyes of others. Similarly, what appears to be self-signaling in private and anonymous settings must represent the workings of a psychology well-adapted to the social incentives of everyday life, in which individuals take advantage of plausible deniability to avoid contributing to public goods, avoid being caught lying and more generally avoid being punished. Overall, this chapter makes the case that our minds are—by necessity—socially adapted and that observed behavior in private or anonymous settings often reflects the workings of that socially adapted psychology, giving the false impression that individuals are playing games with themselves. Since proximate psychological mechanisms (beliefs, preferences, etc.) adapt to the environment in which individuals find

themselves, I concluded this chapter by emphasizing the need for systemic-level interventions for promoting (incentivizing) costly prosocial behavior.

In Chapter 2, I similarly argued that our concern for others, which can materialize through feelings of *warm-glow* or a taste for *fairness*, itself requires further investigation. In line with the perspective adopted in this thesis, I sought to uncover the material and social rewards underlying the expression of our social preferences. As such, I proposed that the theory of social evolution provides a useful theoretical framework for understanding human social preferences. The theory of social evolution predicts that costly cooperative acts need to be (probabilistically) recovered, such that cooperation is expected to provide direct and/or indirect benefits to individuals. Therefore, cooperation (and prosocial behavior, more generally) needs to be incentivized and can be studied using standard economic and game-theoretic tools. I proposed that social emotions such as empathy, anger, guilt, or shame do not exogenously "bias" our decisions in social dilemmas. Instead, they are context-dependent (endogenous) mechanisms that modulate our behavior so as to implement successful strategies in the social domain. I also argued that some quirky features of prosocial behavior, such as "avoiding the ask", "ineffective altruism", or "moral wiggle room" can successfully be understood using this framework. Overall, this chapter makes the case that our (social) preferences and emotions themselves respond to incentives and that a better understanding of how these incentives operate is necessary in order to improve our understanding of human social behavior.

In Chapter 3, I delved into the function of *social identity*, which I defined as the set of beliefs and values that categorize individuals into some subset (or group) of individuals. Individuals are often emotionally attached to their social identity and eager to express it publicly. Researchers have noted that individuals derive *self-esteem* from their group memberships but suffer *psychological costs* from their perceived distance from other group members. Therefore, individuals seek similarity (in terms of preferences, beliefs, etc.) with other group members, but existing research on social identity has not adequately addressed why individuals would have such a preference. In this chapter, I argued that social identity could serve as a signal of trustworthiness in polarized environments. That is, in contexts in which different social groups have associated different beliefs and values that are in conflict with those of other groups, the beliefs and values that individuals decide to adopt can become truthful signals of intention to cooperate with other group members. The incentive to appear as trustworthy can therefore explain why individuals are (emotionally) attached to their social identity and why they seek similarity with other group members. Moreover, viewing social identity as a signal of trustworthi-

ness in polarized environments helps explain several empirical puzzles, such as the malleability and the environment-dependency of social identity or the resistance of social identity to conflicting evidence. This chapter further demonstrates that valuable insights can be garnered by investigating the incentives underlying proximate psychological mechanisms (here, beliefs and values).

In Chapter 4, I aimed to provide theoretical and empirical support to the persuasive account of *positive illusions*. Positive illusions are defined as enhanced beliefs about oneself (one's skills, health, intelligence, etc.) and a variety of theories have been developed to explain how such biased beliefs can remain stable given their potentially high costs. The leading theory, predominantly embraced by psychologists and behavioral economists, suggests that positive illusions provide individuals with *psychological benefits* and improve well-being, counterbalancing the potential costs deriving from sub-optimal decisions. Yet, such a perspective does not explain why adopting positive illusions might feel good. In line with the perspective taken throughout this thesis, the persuasive account of positive illusions suggests that they help individuals persuade others about their (enhanced) qualities, thereby influencing how others behave towards us. Nevertheless, this theory lacks theoretical and empirical support. In this chapter, I showed that the persuasive account of positive illusions is theoretically sound and empirically supported. More specifically, it uniquely predicts that positive illusions will be sensitive to the degree of observability of the trait, the reputational costs of lying, and the ease with which individuals can plausibly deny having self-enhanced. The large empirical literature on positive illusions appears to support these predictions. Following our discussion in Chapter 3, this chapter similarly demonstrates that investigating the function of systematically biased beliefs can help us garner valuable insights into the incentives underlying and maintaining the bias.

In Chapter 5, co-authored with Gisèle Umbhauer, we revisit the Sequential Equilibrium, one of the central equilibrium solution concepts in dynamic games of imperfect/incomplete information, by relaxing the restrictions placed on the players' beliefs at out-of-equilibrium information sets. While the concept of Sequential Equilibrium (SE) requires that all players share the same beliefs about the numerical values of mathematical artifacts used to generate perturbations of strategy profiles, our Individually-Consistent Sequential Equilibrium (ICSE) solution concept allows different perturbation systems for different players. In line with the perspective taken in the other chapters of this thesis, we investigate the strategic nature of beliefs at out-of-equilibrium strategy sets. First, as discussed in Chapter 3, different players might want their beliefs to be as close as possible, particularly so if they be-

long to the same community. This led us to discuss the notion of *distance* between beliefs which can be computed in several ways. Second, the beliefs that the players reveal through their actions might be different from the ones they announce at the start of the game. Third, players might *build* their beliefs at out-of-equilibrium strategy sets to maximize their own payoffs. This suggests that it may be beneficial for players to adopt different beliefs if holding different beliefs can improve their payoffs. These observations led us to reconsider the traditional concept of sequential rationality by additionally requiring that there does not exist unilateral beliefs deviations that are individually-consistent and that can provide more significant payoffs to the player.

Perspectives for Future Work

The work in this thesis sought to investigate the ultimate causes of some behavioral biases documented by economists and psychologists in recent decades. This work was motivated by the idea that behavioral biases do not arise exogenously but can be successfully predicted by appropriately taking into account the social incentives (trustworthiness, reputation, etc.) that individuals face. As such, the present work positions itself in a growing research program that aims at deciphering the incentives underlying our beliefs, preferences, and intuitions (Hoffman & Yoeli 2022). The central insight stemming from this research program is that traditional game-theoretical tools can be successfully applied to understand psychological quirks (or biases) such as motivated reasoning, indirect speech, ineffective altruism, passions, or our sense of justice. The promise is that a better understanding of the incentives shaping our beliefs and preferences can help us better predict and influence behavior, therefore helping overcome some of the most pressing societal issues.

This research program is still in its infancy and there inevitably remains a fertile ground for research questions to be addressed. Concerning the work developed in this thesis, I can see several avenues to extend and improve on it. In Chapter 1, I suggested that feelings can become incentives (or rewards) only in fitness-irrelevant domains, such as entertainment. The argument was that letting feelings dictate behavior in fitness-relevant domains, such as one's health, skills, or mating prospects would be too costly and, therefore, unlikely to be evolutionary stable. Yet, what exactly distinguishes fitness-relevant from fitness-irrelevant domains? What are the relevant benefits and costs that need to be taken into account to discriminate between both domains? More generally, under what conditions can feelings become incentives? Additionally, more work is needed to better understand the functioning of human learning mechanisms. I have suggested that the *Reinforcement Leaning*

and *Social Learning* frameworks capture a large class of learning problems, but it is not yet clear how fine-tuned these mechanisms are. Can we expect humans to optimally adapt their behavior to every contingency they have encountered or are learning mechanisms bound to respond to the "average" contingency? Finally, I have suggested that the development of institutions that promote and incentivize cooperation is crucial in order to tackle the most pressing societal issues. Yet, it is unclear how such institutions might be "socially-engineered" from scratch. Recent work has shown that "too strict" social norms can backfire, increasing rather than decreasing rule-violations (Aycinena et al. 2022). Therefore, more work is required to better understand how to successfully design institutions incentivizing prosocial behavior.

In Chapter 2, I have suggested, based on the logic of evolutionary theory, that a trait (or behavior) is expected to be adaptive only in those circumstances in which it was selected for. This proposition predicts that behavior might not be adaptive in unfamiliar environments such as the laboratory. I have argued that higher-than-expected cooperation in the laboratory might result from such a mismatch, wherein an adaptive behavior in one environment spillovers in another. Implicit in this argument is the idea that there must exist some cues in the laboratory that remind individuals of out-of-the-laboratory (everyday life) settings. An important avenue for future research would be to specify whether there exist aspects of the environment that can reliably predict whether a behavior will spill over from one setting to another. That is, can we predict, based on certain similarities or differences between environments, whether spillovers will occur from one setting to the other? This prediction would provide solid theoretical foundations for spillover effects and would undoubtedly improve our understanding of human behavior. Additionally, I have suggested that cultural/institutional evolution and behavioral ecology are powerful frameworks for understanding the wide variation in the expression of social preferences. Yet, more research is needed to understand how these two mechanisms interact. Are there circumstances under which we can expect one mechanism to be more important than the other? Finally, by focusing on prosocial behavior as a strategy, I have not considered research on personality traits such as agreeableness. While stable exposure to mechanisms incentivizing cooperation might be confused with a stable individual-level predisposition, future research could clarify to what extent prosociality can be considered an individual-level (stable) trait.

In Chapter 3, I suggested that social identity can become a signal of trustworthiness in polarized environments. Yet, the theoretical part of this paper takes this fact as given. A natural extension of the present work would be to integrate the (verbal)

argument that social identity can signal trustworthiness (based on Loury (1994)'s work) with the formalization of the process of social identity adoption presented in this paper. While Golman (2022) has started to work in that direction, he takes individual values as given and is not concerned with trustworthiness. Following our approach, social identity would endogenously become a signal of trustworthiness and the choice of social identity would remain a function of social incentives (cooperation opportunities). Future work could also more thoroughly describe the conditions under which social identity can become a truthful signal of trustworthiness as well as the conditions under which social identity might have other functions, such as coordination with similar others (Smaldino 2019). Finally, more work is needed to clearly delineate the role of information and psychology in understanding which social identity individuals decide to adopt. Researchers have suggested that some available empirical evidence is consistent with a Bayesian account of political belief formation (Tappin et al. 2020) or that particular psychological mechanisms can influence information-processing and bias the process of political belief formation (Funk et al. 2013). Future work could assess the relative importance of these mechanisms with respect to the signaling account developed in this thesis and further test the empirical significance of the signaling function of social identity.

In Chapter 4, while I assumed that all S types benefit in the same way from interacting with R in the "Partner Choice" game, one might wonder how the equilibrium results would change if "Lower" types benefited more than "Higher" types. Might R be more prudent and less easily persuaded by positive illusions? Similarly, how might the results change if the reputational costs of lying were type-dependent, with, for instance, "Higher" types valuing their reputation more? Finally, while dynamics have been left out of the present theoretical analysis, it would be interesting to know how repeated play between S and R might affect the stability of positive illusions. In the "Partner Choice" game, can repeated interactions (and therefore repeated feedback about S's type) still allow some room for positive illusions to persuade at equilibrium? In the "Community" game, can repeated play reduce the Receiver's uncertainty about each other's "punishment threshold" and therefore improve coordination? If a lie is punished, could a different lie maintaining plausible deniability replace it? Moreover, with respect to the collective punishment of lying, several interesting questions are raised. Under what conditions is it beneficial to punish a liar alone? Alternatively, when is it risky to punish if others do not follow? What kind of information-structure favors common knowledge of lying among different players? Answering these questions will undoubtedly improve our understanding of positive illusions and might refine the arguments exposed in this

chapter. More generally, one important question for future research would be to precisely determine the factors influencing plausible deniability. While I have suggested that plausible deniability can prevent coordinated punishment and stabilize positive illusions at equilibrium, future work should investigate what is plausibly deniable and what is not. An extensive theoretical and empirical treatment of plausible deniability promises to illuminate a variety of phenomena, such as contributions to public goods (see Chapter 2), ethical behavior, or political discourse.

In Chapter 5, we suggested that the Sequential Equilibrium (SE) requirement that players share the same beliefs about out-of-equilibrium actions was too strong and proposed to weaken this condition in our Individually-Consistent Sequential Equilibrium (ICSE). We have shown that it can be advantageous for players to hold different beliefs if such beliefs can allow them to obtain greater payoffs. An avenue for future research would be to further investigate the conditions under which heterogeneous beliefs might arise at equilibrium, and their relationship with the payoff-structure. Also, how do opposing incentives influence the players' equilibrium beliefs? As a matter of example, how does the incentive to adopt beliefs close to the ones adopted by other community members (to appear as trustworthy) relate to the incentive to adopt accurate beliefs about the state of the world or to the incentive to adopt *different* beliefs so as to influence other players' decisions? More generally, future research could explore in more detail the consequences of adding beliefs to the players' strategy set.

To conclude, I want to note that across all chapters of this thesis, I have implicitly (if not explicitly) assumed that there are no constraints preventing individuals to behave optimally or preventing psychological proximate mechanisms (beliefs, preferences, etc.) to adapt optimally to the incentives in the environment. I hope that the present work has convinced the reader that such a perspective can bring interesting insights and help us better understand human (social) behavior. Yet, in reality, there certainly exist constraints that prevent individuals to behave optimally in the circumstances in which they find themselves. An important avenue for future research would be to clarify exactly what these constraints are and under what circumstances we can expect them to be binding. For instance, are there general limits on learning that can prevent individuals from discriminating between contexts or general computational limits that can prevent individuals from responding to relevant incentives? How does the structure of social networks, individual rationality or inertia influence the outcome of learning processes? Are there domains in which we can expect "genuine cognitive biases" to persist? Are there fundamental constraints, due to the way our brains have evolved, that might prevent individuals from success-

fully adapting and responding to the contingencies which they face? Also, what are the evaluation criteria that individuals make use of in order to determine whether a trait (or behavior) is useful or not (Singh 2020) and how do these criteria respond to real-world feedback? I believe that a greater understanding of how evolution has shaped human psychological mechanisms—for what purposes and with what constraints—is needed to better predict when human behavior can be expected to be adapted to its environment.

Conclusion générale

L'objectif principal de cette thèse a été de réévaluer le tournant "comportemental" en économie et en théorie des jeux. Ce tournant "comportemental", inspiré par la recherche en psychologie, a été motivé par l'observation que les individus ne se comportent souvent pas conformément aux prédictions des modèles économiques et de théorie des jeux classique basés sur la maximisation de l'intérêt personnel et la rationalité Bayésienne. Au contraire, les individus adoptent souvent des croyances qui ne sont pas justifiées par l'évidence, prennent en compte le bien-être des autres dans les dilemmes sociaux et, plus généralement, semblent se fier davantage aux émotions qu'à la "rationalité" lorsqu'ils prennent des décisions. Dans cette thèse, j'ai proposé que les déviations par rapport aux prédictions des modèles standards peuvent souvent être expliquées à l'aide d'outils économiques et de la théorie des jeux classique, à condition que les chercheurs tiennent compte des incitations sociales auxquelles les individus sont confrontés. L'argument que j'ai essayé d'avancer est que nos mécanismes psychologiques et émotionnels ne biaisent pas nos décisions de manière exogène mais qu'ils répondent eux-mêmes à des incitations qui peuvent ne pas être directement apparentes. Par conséquent, les outils basés sur la rationalité individuelle et la maximisation de l'intérêt personnel peuvent être appliqués de manière fructueuse afin de comprendre comment ces mécanismes répondent aux incitations.

Dans le Chapitre 1, j'ai soutenu l'idée que le *sentiment* de vouloir améliorer son image de soi et le *besoin* subséquent de s'auto-signaler nécessitent eux-mêmes une explication plus ultime, puisque les deux (considérés individuellement) sont purement inconséquents. Étant donné que les mécanismes d'apprentissage et les mécanismes émotionnels répondent aux récompenses et aux punitions dans notre environnement, le sentiment positif associé à une image positive de soi doit provenir de bénéfices matériels ou sociaux associés à une image positive. À l'inverse, le sentiment négatif associé à un comportement incohérent ou à un comportement socialement désapprouvé doit découler des coûts matériels ou sociaux associés à ces comportements. Cela m'a amené à conclure que l'image de soi désirée doit représenter l'image sociale désirée, dans le sens où les individus sont ultimement motivés à améliorer et maintenir leur réputation aux yeux des autres. De même, ce qui semble être un auto-signal dans des contextes privés et anonymes doit représenter le fonctionnement d'une psychologie bien adaptée aux incitations sociales de la vie quotidienne, dans laquelle les individus tirent parti de facteurs contextuels afin d'éviter de contribuer aux biens publics, d'être pris en train de mentir et, plus généralement, d'être

punis. Dans l'ensemble, ce chapitre montre que nos esprits sont—par nécessité—socialement adaptés et que le comportement observé dans des contextes privés ou anonymes reflète souvent le fonctionnement de cette psychologie socialement adaptée, donnant la fausse impression que les individus jouent à des jeux avec eux-mêmes. Étant donné que les mécanismes psychologiques proximaux (croyances, préférences, etc.) s'adaptent à l'environnement dans lequel les individus se trouvent, j'ai conclu ce chapitre en soulignant la nécessité d'interventions au niveau systémique pour inciter (encourager) des comportements prosociaux coûteux.

Dans le Chapitre 2, j'ai également soutenu que notre souci pour le bien-être des autres, qui peut se matérialiser par des sentiments de *warm-glow* ou un goût pour l'*équité*, nécessite lui-même une analyse plus approfondie. Conformément à la perspective adoptée dans cette thèse, j'ai cherché à découvrir les récompenses matérielles et sociales sous-jacentes à l'expression de nos préférences sociales. À ce titre, j'ai proposé que la théorie de l'évolution sociale fournit un cadre théorique utile permettant de mieux comprendre les préférences sociales. La théorie de l'évolution sociale prédit que les actes de coopération coûteux doivent être bénéfiques (de manière probabiliste), de sorte que la coopération est censée apporter des bénéfices directs et/ou indirects aux individus. Par conséquent, la coopération (et plus généralement le comportement prosocial) doit être incitée et peut être étudiée à l'aide d'outils économiques et de la théorie des jeux classique. J'ai proposé que les émotions sociales telles que l'empathie, la colère, la culpabilité ou la honte ne "biaisent" pas de manière exogène nos décisions dans les dilemmes sociaux. Il s'agit plutôt de mécanismes (endogènes) dépendant du contexte qui modulent notre comportement de manière à mettre en œuvre des stratégies efficaces dans le domaine social. J'ai également fait valoir que certaines caractéristiques étonnantes du comportement prosocial (comme l'altruisme inefficace) peuvent être comprises à travers ce cadre théorique. Dans l'ensemble, ce chapitre montre que nos préférences sociales et nos émotions répondent elles-mêmes à des incitations et qu'une meilleure compréhension du fonctionnement de ces incitations est nécessaire afin d'améliorer notre compréhension du comportement social humain.

Dans le Chapitre 3, j'ai analysé la fonction de l'*identité sociale* que j'ai défini comme représentant l'ensemble des croyances et des valeurs qui classent les individus dans un certain sous-ensemble (ou groupe) d'individus. Les individus sont souvent émotionnellement attachés à leur identité sociale et désireux de l'exprimer publiquement. Les chercheurs ont remarqué que les individus retirent de l'*estime de soi* de leur appartenance à un groupe mais souffrent de *coûts psychologiques* dûs à leur distance perçue par rapport aux autres membres du groupe. Par conséquent,

les individus recherchent la similarité (en termes de préférences, de croyances, etc.) avec les autres membres du groupe, mais les recherches existantes sur l'identité sociale n'ont pas abordé de manière adéquate les raisons pour lesquelles les individus auraient une telle préférence. Dans ce chapitre, j'ai soutenu que l'identité sociale peut servir de signal de fiabilité dans des environnements polarisés. En d'autres termes, dans des contextes où différents groupes sociaux ont associé des croyances et des valeurs qui sont en conflit avec celles d'autres groupes, les croyances et les valeurs que les individus décident d'adopter peuvent devenir des signaux d'intention de coopérer avec les autres membres de leur groupe. L'incitation à paraître comme digne de confiance peut donc expliquer pourquoi les individus sont (émotionnellement) attachés à leur identité sociale et pourquoi ils recherchent la similarité avec les autres membres du groupe. De plus, considérer l'identité sociale comme un signal de fiabilité dans des environnements polarisés permet d'expliquer plusieurs puzzles empiriques tels que la malléabilité de l'identité sociale et la dépendance de l'identité sociale vis-à-vis de l'environnement ou la résistance de l'identité sociale aux informations contradictoires. Ce chapitre démontre que l'on peut obtenir une différente perspective sur d'importants phénomènes individuels et sociaux en étudiant les incitations sous-jacentes aux mécanismes psychologiques proximaux (ici les croyances et les valeurs).

Dans le Chapitre 4, j'ai cherché à apporter une base théorique et empirique à l'approche "persuasive" des *illusions positives*. Les illusions positives sont définies comme des croyances sur soi-même (ses compétences, sa santé, son intelligence, etc.) positivement biaisées, et diverses théories ont été développées afin d'expliquer comment de telles croyances biaisées peuvent rester stables compte tenu de leurs coûts potentiellement élevés. La principale théorie, à laquelle adhèrent notamment les psychologues et les économistes comportementaux, suggère que les illusions positives procurent aux individus des *bénéfices psychologiques* et améliorent le bien-être, contrebalançant ainsi les coûts potentiels découlant de décisions sous-optimales. Cependant, une telle perspective n'explique pas pourquoi l'adoption d'illusions positives peut procurer un sentiment de bien-être. Conformément à la perspective adoptée dans le reste de cette thèse, l'approche "persuasive" des illusions positives suggère que les illusions positives aident les individus à persuader les autres de leurs qualités (améliorées), influençant ainsi le comportement des autres à leur égard. Néanmoins, cette théorie manque de soutien théorique et empirique. Dans ce chapitre, j'ai montré que l'approche "persuasive" des illusions positives est théoriquement solide et empiriquement soutenue. Plus précisément, cette approche prédit de manière unique que les illusions positives seront sensibles au degré d'observabilité du trait,

aux coûts réputationnels liés au mensonge et à la facilité avec laquelle les individus peuvent plausiblement nier avoir menti. L'importante littérature empirique sur les illusions positives semble soutenir ces prédictions. Suite à notre discussion dans le Chapitre 3, ce chapitre démontre également que l'étude de la fonction des croyances systématiquement biaisées peut nous offrir une nouvelle perspective sur les incitations qui sous-tendent et maintiennent le biais.

Dans le Chapitre 5, co-écrit avec Gisèle Umbhauer, nous revisitons l'équilibre séquentiel, l'un des concepts d'équilibre centraux dans les jeux dynamiques d'information imparfaite/incomplète, en réduisant les restrictions placées sur les croyances des joueurs aux ensembles d'information hors équilibre. Alors que le concept d'équilibre séquentiel (ES) exige que tous les joueurs partagent les mêmes croyances sur les valeurs numériques d'artefacts mathématiques utilisés pour générer des perturbations des profils de stratégies, notre concept de solution d'équilibre séquentiel individuellement-cohérent (ESIC) permet différents systèmes de perturbation pour différents joueurs. En accord avec la perspective adoptée dans les autres chapitres de cette thèse, nous étudions la nature stratégique des croyances aux ensembles d'information hors équilibre. Tout d'abord, comme discuté dans le Chapitre 3, différents joueurs peuvent souhaiter que leurs croyances soient aussi proches que possible, particulièrement s'ils appartiennent à une même communauté. Cela nous a conduit à discuter une notion de *distance* entre les croyances qui peut être calculée de plusieurs façons. Deuxièmement, les croyances que les joueurs révèlent par leurs actions peuvent être différentes de celles qu'ils annoncent au début du jeu. Troisièmement, les joueurs peuvent *construire* leurs croyances à des ensembles d'information hors équilibre de telle sorte à maximiser leurs propres gains. Ceci suggère qu'il peut être bénéfique pour les joueurs d'adopter des croyances différentes si celles-ci leur permettent d'améliorer leurs gains. Ces observations nous ont amené à reconsidérer le concept traditionnel de rationalité séquentielle, en exigeant en plus qu'il n'existe pas de déviations unilatérales des croyances qui soient individuellement-cohérentes et qui puissent fournir des gains plus importants au joueur.

Perspectives pour des recherches futures

Le travail de cette thèse a visé à étudier les causes ultimes de certains biais comportementaux documentés par les économistes et les psychologues au cours des dernières décennies. Ce travail a été motivé par l'idée que les biais comportementaux ne surviennent pas de façon exogène mais peuvent être prédits avec succès en prenant en compte de manière appropriée les incitations sociales (fiabilité, réputation, coopération, etc.) auxquelles les individus sont confrontés. En tant que tel, le présent

travail se positionne dans un programme de recherche naissant qui vise à analyser les incitations sous-jacentes à nos croyances, nos préférences et nos intuitions (Hoffman & Yoeli 2022). L'idée principale de ce programme de recherche est que les outils traditionnels de la théorie des jeux peuvent être appliqués avec succès pour comprendre les "biais" psychologiques tels que le raisonnement motivé, le discours indirect, l'altruisme inefficace, les passions ou notre sens de la justice. La promesse est qu'une meilleure compréhension des incitations qui façonnent nos croyances et nos préférences peut nous aider à mieux prédire et influencer le comportement et donc à surmonter certains des enjeux sociétaux les plus urgents.

Ce programme de recherche n'en est qu'à ses débuts et il reste inévitablement un terrain fertile de questions de recherche à aborder. En ce qui concerne le travail développé dans cette thèse, je vois plusieurs pistes pour l'étendre et l'améliorer. Dans le Chapitre 1, j'ai suggéré que les sentiments peuvent devenir des incitations (ou des récompenses) uniquement dans des domaines inconséquents en termes de survie et de reproduction, comme le divertissement. L'argument était que laisser les sentiments dicter le comportement dans des domaines conséquents, tels que la santé, serait trop coûteux et donc peu susceptible d'être stable sur le plan évolutif. Néanmoins, que distingue exactement les domaines conséquents pour la survie et la reproduction de ceux qui ne le sont pas ? Quels sont les bénéfices et les coûts à prendre en compte afin de faire la distinction entre ces deux domaines ? Plus généralement, sous quelles conditions les sentiments peuvent-ils devenir des incitations ? En outre, des recherches supplémentaires sont nécessaires afin de mieux comprendre le fonctionnement des mécanismes d'apprentissage humains. J'ai suggéré que l'*apprentissage par renforcement* et l'*apprentissage social* capturent une grande classe de problèmes d'apprentissage, mais il n'est pas encore clair dans quelle mesure ces mécanismes sont optimalement réglés. Peut-on s'attendre à ce que les humains adaptent de façon optimale leur comportement à toutes les éventualités qu'ils ont rencontrées ou est-ce que les mécanismes d'apprentissage sont-ils tenus de répondre à l'éventualité "moyenne" ? Enfin, j'ai suggéré que le développement d'institutions qui favorisent et encouragent la coopération est crucial pour s'attaquer aux problèmes sociétaux les plus urgents. Pourtant, la manière dont de telles institutions peuvent être "socialement façonnées" n'est pas claire. Des travaux récents ont montré que des normes sociales "trop strictes" peuvent être contre-productives, augmentant plutôt que diminuant les violations des règles (Aycinena et al. 2022). Par conséquent, des travaux supplémentaires sont nécessaires afin de mieux comprendre comment concevoir avec succès des institutions incitant le comportement prosocial coûteux.

Dans le Chapitre 2, j'ai suggéré, en me basant sur la logique de la théorie de l'évolution, qu'un trait (ou un comportement) est censé être adaptatif uniquement dans les circonstances dans lesquelles il a été sélectionné. Cela permet de prédire qu'un comportement peut ne pas être adaptatif dans des environnements non familiers tels que le laboratoire. J'ai soutenu que la coopération plus importante que prévu en laboratoire pourrait résulter d'une telle inadéquation, dans laquelle un comportement adaptatif dans un environnement est "transféré" dans un autre environnement. L'idée implicite dans cet argument est qu'il doit exister des indications au sein du laboratoire qui rappellent aux individus certaines situations hors laboratoire (de la vie quotidienne). Une importante piste de recherche future consisterait à préciser s'il existe des aspects de l'environnement qui permettent de prédire de manière fiable si un comportement va être "transféré" d'un environnement à un autre. En d'autres termes, pouvons-nous prédire, sur la base de certaines similitudes ou différences entre les environnements, si des "transferts" se produiront d'un cadre à l'autre ? Cela fournirait des bases théoriques solides pour les effets de "transfert" et améliorerait sans aucun doute notre compréhension du comportement humain. En outre, j'ai suggéré que l'évolution culturelle/institutionnelle et l'écologie comportementale sont des cadres théoriques importants pour comprendre la grande variation dans l'expression des préférences sociales. Pourtant, des recherches supplémentaires sont nécessaires afin de comprendre comment ces deux mécanismes interagissent. Y a-t-il des circonstances dans lesquelles nous pouvons nous attendre à ce qu'un mécanisme soit plus important que l'autre ? Enfin, en me concentrant sur le comportement prosocial en tant que stratégie, je n'ai pas pris en compte les recherches sur les traits de personnalité tels que l'agréabilité. Si une exposition stable à des mécanismes incitant à la coopération peut être confondue avec une prédisposition stable au niveau individuel, des recherches futures pourraient clarifier dans quelle mesure la prosocialité peut être considérée comme un trait (stable) au niveau individuel ou une stratégie (comme suggéré dans cette thèse).

Dans le Chapitre 3, j'ai suggéré que l'identité sociale peut devenir un signal de fiabilité dans les environnements polarisés. Pourtant, la partie théorique de cet article prend ce fait pour acquis. Une extension naturelle du présent travail serait d'intégrer l'argument (verbal) selon lequel l'identité sociale peut signaler la fiabilité (basé sur le travail de Loury 1994) avec la formalisation du processus d'adoption de l'identité sociale présenté dans ce chapitre. Bien que Golman (2022) ait commencé à travailler dans cette direction, il prend les valeurs individuelles comme données et ne s'intéresse pas à la confiance. En suivant notre approche, l'identité sociale deviendrait de manière endogène un signal de fiabilité et le choix de l'identité so-

ciale resterait une fonction des incitations sociales (opportunités de coopération). Les travaux futurs pourraient également décrire plus en détail les conditions sous lesquelles l'identité sociale peut devenir un signal de fiabilité, ainsi que les conditions sous lesquelles l'identité sociale pourrait avoir d'autres fonctions, telles que la coordination avec d'autres personnes similaires (Smaldino 2019). Enfin, des travaux supplémentaires sont nécessaires pour délimiter clairement le rôle de l'information et de la psychologie dans la compréhension de l'identité sociale que les individus décident d'adopter. Des chercheurs ont suggéré que certains travaux empiriques existants sont compatibles avec une approche Bayésienne de la formation des croyances politiques (Tappin et al. 2020) ou que certains mécanismes psychologiques peuvent influencer le traitement de l'information et biaiser le processus de formation des croyances politiques (Funk et al. 2013). Des travaux futurs pourraient évaluer l'importance relative de ces mécanismes par rapport à l'approche stratégique développée dans cette thèse et tester davantage la signification empirique de la valeur de signal de l'identité sociale.

Dans le Chapitre 4, bien que j'aie supposé que tous les différents types d'émetteurs S bénéficient de la même manière de l'interaction avec le receveur R dans le jeu du "choix du partenaire", on peut se demander comment les résultats à l'équilibre changeraient si les types "inférieurs" bénéficiaient davantage que les types "supérieurs". Est-ce que R serait plus prudent et moins facilement persuadé par des illusions positives ? De même, comment les résultats pourraient-ils changer si les coûts réputationnels du mensonge dépendaient du type, avec, par exemple, des types "supérieurs" valorisant davantage leur réputation ? Enfin, bien que la dynamique ait été laissée de côté dans la présente analyse théorique, il serait intéressant de savoir comment des interactions répétées entre S et R pourraient affecter la stabilité des illusions positives. Dans le jeu du "choix du partenaire", les interactions répétées permettent-elles aux illusions positives de continuer à persuader à l'équilibre ? Dans le jeu de "communauté", les interactions répétées peuvent-elles réduire l'incertitude des receveurs quant au "seuil de punition" de l'autre receveur et donc améliorer la coordination ? Si un mensonge est puni, un autre mensonge permettant de maintenir le déni plausible pourrait-il le remplacer ? De plus, en ce qui concerne la punition collective du mensonge, plusieurs questions intéressantes sont soulevées. Sous quelles conditions est-il bénéfique de punir un menteur seul ? Alternativement, quand est-il risqué de punir si les autres ne suivent pas ? Quel type de structure d'information favorise une connaissance commune du mensonge parmi les différents acteurs ? La réponse à ces questions améliorerait sans aucun doute notre compréhension des illusions positives et pourrait affiner les arguments exposés

dans ce chapitre. Plus généralement, une question importante pour les recherches futures serait de déterminer précisément les facteurs influençant le déni plausible. Bien que j'aie suggéré que le déni plausible peut empêcher la punition coordonnée et donc stabiliser les illusions positives à l'équilibre, les travaux futurs devraient étudier plus en détail ce qui est déniable de façon plausible et ce qui ne l'est pas. Un traitement théorique et empirique approfondi du déni plausible promet d'éclairer une variété de phénomènes tels que les contributions aux biens publics (voir Chapitre 2), le comportement éthique ou le discours politique.

Dans le Chapitre 5, nous avons suggéré que la condition de l'équilibre séquentiel (ES) selon laquelle les joueurs doivent partager les mêmes croyances sur les actions hors équilibre était trop forte, et nous avons proposé de modifié cette condition dans notre équilibre séquentiel individuellement-cohérent (ESIC). Nous avons montré qu'il peut être avantageux pour les joueurs d'avoir des croyances différentes si ces croyances peuvent leur permettre d'obtenir de meilleurs gains. Une piste de recherche future consisterait à étudier plus avant les conditions sous lesquelles des croyances hétérogènes peuvent apparaître à l'équilibre, et leur relation avec la structure des gains. De même, comment des incitations opposées influencent-elles les croyances des joueurs à l'équilibre ? Par exemple, comment l'incitation à adopter des croyances proches de celles adoptées par les autres membres de la communauté (pour apparaître comme digne de confiance) est-elle liée à l'incitation à adopter des croyances exactes sur l'état du monde ou à l'incitation à adopter des croyances *différentes* afin d'influencer les décisions des autres joueurs ? Plus généralement, les recherches futures pourraient explorer plus en détail les conséquences de l'ajout des croyances à l'ensemble des stratégies des joueurs.

Pour conclure, je tiens à noter que dans tous les chapitres de cette thèse, j'ai implicitement (sinon explicitement) supposé qu'il n'y a pas de contraintes empêchant les individus de se comporter de façon optimale ou empêchant les mécanismes psychologiques proximaux (croyances, préférences, etc.) de s'adapter de façon optimale aux incitations dans l'environnement. J'espère que le présent travail a convaincu le lecteur qu'une telle perspective peut éclairer certains phénomènes et nous aider à mieux comprendre le comportement (social) humain. Néanmoins, dans la réalité, il existe certainement des contraintes qui empêchent les individus de se comporter de façon optimale dans les circonstances dans lesquelles ils se trouvent. Une importante piste de recherche future consisterait à clarifier la nature exacte de ces contraintes et les circonstances dans lesquelles nous pouvons nous attendre à ce qu'elles soient contraignantes. Par exemple, existe-t-il des limites générales à l'apprentissage qui peuvent empêcher les individus de discriminer entre différents contextes ou des lim-

ites générales de calcul ou de traitement de l'information qui peuvent empêcher les individus de répondre aux incitations ? Comment la structure des réseaux sociaux, la rationalité ou l'inertie influencent-elles le résultat des processus d'apprentissage ? Existe-t-il des domaines dans lesquels nous pouvons nous attendre à ce que des "biais cognitifs" persistent ? Existe-t-il des contraintes fondamentales, dues à l'évolution de nos cerveaux, qui pourraient empêcher les individus de s'adapter et de répondre avec succès aux contingences auxquelles ils sont confrontés ? Aussi, quels sont les critères d'évaluation que les individus utilisent afin de déterminer si un trait (ou un comportement) est utile ou non (Singh 2020) et comment ces critères réagissent-ils au feedback du monde réel ? Je pense qu'une meilleure compréhension de la manière dont l'évolution a façonné les mécanismes psychologiques humains—à quelles fins et avec quelles contraintes—est nécessaire afin de mieux prédire les circonstances dans lesquelles le comportement humain peut être adapté à son environnement.

Bibliography

Adolphs, R. & Anderson, D. J. (2018), *The neuroscience of emotion: A new synthesis*, Princeton University Press.

Agadjanian, A. & Lacy, D. (2021), 'Changing votes, changing identities? racial fluidity and vote switching in the 2012–2016 us presidential elections', *Public Opinion Quarterly* **85**(3), 737–752.

Akerlof, G. A. & Kranton, R. E. (2000), 'Economics and identity', *The Quarterly Journal of Economics* **115**(3), 715–753.

Alicke, M. D. (1985), 'Global self-evaluation as determined by the desirability and controllability of trait adjectives.', *Journal of Personality and Social Psychology* **49**(6), 1621.

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J. & Vredenburg, D. S. (1995), 'Personal contact, individuation, and the better-than-average effect.', *Journal of Personality and Social Psychology* **68**(5), 804.

Allison, S. T., Messick, D. M. & Goethals, G. R. (1989), 'On being better but not smarter than others: The muhammad ali effect', *Social Cognition* **7**(3), 275–295.

Alpizar, F., Carlsson, F. & Johansson-Stenman, O. (2008), 'Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in costa rica', *Journal of Public Economics* **92**(5-6), 1047–1060.

Alvard, M. (2009), 'Kinship and cooperation', *Human Nature* **20**(4), 394.

Anderson, C., Brion, S., Moore, D. A. & Kennedy, J. A. (2012), 'A status-enhancement account of overconfidence.', *Journal of Personality and Social Psychology* **103**(4), 718.

Andreoni, J. (1990), 'Impure altruism and donations to public goods: A theory of warm-glow giving', *The Economic Journal* **100**(401), 464–477.

Andreoni, J. & Miller, J. H. (1993), 'Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence', *The Economic Journal* **103**(418), 570–585.

Andreoni, J., Rao, J. M. & Trachtman, H. (2017), 'Avoiding the ask: A field experiment on altruism, empathy, and charitable giving', *Journal of Political Economy* **125**(3), 625–653.

Anthony, D. B., Holmes, J. G. & Wood, J. V. (2007), 'Social acceptance and self-esteem: tuning the sociometer to interpersonal value.', *Journal of Personality and Social Psychology* **92**(6), 1024.

Aspinwall, L. G. & Tedeschi, R. G. (2010), 'The value of positive psychology for health psychology: Progress and pitfalls in examining the relation of positive phenomena to health', *Annals of Behavioral Medicine* **39**(1), 4–15.

Aumann, R. J. (1976), 'Agreeing to disagree', *The Annals of Statistics* **4**(6), 1236–1239.

Aumann, R. J. (2019), 'A synthesis of behavioural and mainstream economics', *Nature Human Behaviour* **3**(7), 666–670.

Axelrod, R. & Hamilton, W. (1981), 'The evolution of cooperation', *Science* **211**(4489), 1390–1396.

Aycinena, D., Rentschler, L., Beranek, B. & Schulz, J. F. (2022), 'Social norms and dishonesty across societies', *Proceedings of the National Academy of Sciences* **119**(31), e2120138119.

Ayres, I., Raseman, S. & Shih, A. (2013), 'Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage', *The Journal of Law, Economics, and Organization* **29**(5), 992–1022.

Baggini, J. (2011), *The ego trick*, Granta Books.

Balbuzanov, I. (2019), 'Lies and consequences', *International Journal of Game Theory* **48**(4), 1203–1240.

Barber, B. M. & Odean, T. (2001), 'Boys will be boys: Gender, overconfidence, and common stock investment', *The Quarterly Journal of Economics* **116**(1), 261–292.

Barber, M. & Pope, J. C. (2019), 'Does party trump ideology? Disentangling party and ideology in America', *American Political Science Review* **113**(1), 38–54.

Barreda-Tarrazona, I., Jaramillo-Gutiérrez, A., Pavan, M. & Sabater-Grande, G. (2017), 'Individual characteristics vs. experience: An experimental study on cooperation in prisoner's dilemma', *Frontiers in Psychology* **8**, 596.

Barrett, H. C. (2014), *The shape of thought: How mental adaptations evolve*, Oxford University Press.

Barrett, J. L. (2000), 'Exploring the natural foundations of religion', *Trends in Cognitive Sciences* **4**(1), 29–34.

Barrett, L. F. (2017), *How emotions are made: The secret life of the brain*, Pan Macmillan.

Bašić, Z. & Quercia, S. (2022), 'The influence of self and social image concerns on lying', *Games and Economic Behavior* **133**, 162–169.

Baumard, N., André, J.-B. & Sperber, D. (2013), 'A mutualistic approach to morality: The evolution of fairness by partner choice', *Behavioral and Brain Sciences* **36**(1), 59–78.

Baumeister, R. F. (1989), 'The optimal margin of illusion', *Journal of Social and Clinical Psychology* **8**(2), 176–189.

Baumeister, R. F. (2019), 'Stalking the true self through the jungles of authenticity: Problems, contradictions, inconsistencies, disturbing findings—and a possible way forward', *Review of General Psychology* **23**(1), 143–154.

Baumeister, R. F. (2022), *The self explained: Why and how we become who we are*, Guilford Publications.

Baumeister, R. F., Campbell, J. D., Krueger, J. I. & Vohs, K. D. (2003), 'Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles?', *Psychological Science in the Public Interest* **4**(1), 1–44.

Baumeister, R. F., Heatherton, T. F. & Tice, D. M. (1993), 'When ego threats lead to self-regulation failure: Negative consequences of high self-esteem.', *Journal of Personality and Social Psychology* **64**(1), 141.

Bear, A. & Rand, D. G. (2016), 'Intuition, deliberation, and the evolution of cooperation', *Proceedings of the National Academy of Sciences* **113**(4), 936–941.

Beltran, D. G., Shiota, M. N. & Aktipis, A. (2022), 'On the proximate and ultimate functions of the social emotions with regard to cooperation', *To appear in L. Al-Shawaf and T. K. Shackelford (Eds.), The Oxford Handbook of Evolution and the Emotions* .

Bénabou, R. (2015), 'The economics of motivated beliefs', *Revue d'Economie Politique* **125**(5), 665–685.

Bénabou, R. & Tirole, J. (2002), 'Self-confidence and personal motivation', *The Quarterly Journal of Economics* **117**(3), 871–915.

Bénabou, R. & Tirole, J. (2004), 'Willpower and personal rules', *Journal of Political Economy* **112**(4), 848–886.

Bénabou, R. & Tirole, J. (2011), 'Identity, morals, and taboos: Beliefs as assets', *The Quarterly Journal of Economics* **126**(2), 805–855.

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

Berg, J., Dickhaut, J. & McCabe, K. (1995), 'Trust, reciprocity, and social history', *Games and Economic Behavior* **10**(1), 122–142.

Billing, J. & Sherman, P. W. (1998), 'Antimicrobial functions of spices: why some like it hot', *The Quarterly Review of Biology* **73**(1), 3–49.

Birch, J. & Okasha, S. (2015), 'Kin selection and its critics', *BioScience* **65**(1), 22–32.

Birch, S. A., Vauthier, S. A. & Bloom, P. (2008), 'Three-and four-year-olds spontaneously use others' past performance to guide their learning', *Cognition* **107**(3), 1018–1034.

Bodner, R. & Prelec, D. (2003), Self-signaling and diagnostic utility in everyday decision making, *in* '*I. Brocas and J. D. Carrillo (Eds.), The Psychology of Economic Decisions.*', Vol. 1, Oxford: Oxford University Press, pp. 105—123.

Bolger, D., Thomson Jr, R. A. & Ecklund, E. H. (2019), 'Selection versus socialization? Interrogating the sources of secularity in global science', *Sociological Perspectives* **62**(4), 518–537.

Bonanno, G. (2013), 'Agm-consistency and perfect bayesian equilibrium. part i: definition and properties', *International Journal of Game Theory* **42**(3), 567–592.

Bonanno, G. (2016), 'Agm-consistency and perfect bayesian equilibrium. part ii: from pbe to sequential equilibrium', *International Journal of Game Theory* **45**(4), 1071–1094.

Bonica, A. (2014), 'Mapping the ideological marketplace', *American Journal of Political Science* **58**(2), 367–386.

Bowles, S. (2006), 'Group competition, reproductive leveling, and the evolution of human altruism', *Science* **314**(5805), 1569–1572.

Bowles, S. & Gintis, H. (2004), 'The evolution of strong reciprocity: cooperation in heterogeneous populations', *Theoretical Population Biology* **65**(1), 17–28.

Boyd, R. (2017), *A different kind of animal: How culture transformed our species*, Vol. 46, Princeton University Press.

Boyd, R., Gintis, H. & Bowles, S. (2010), 'Coordinated punishment of defectors sustains cooperation and can proliferate when rare', *Science* **328**(5978), 617–620.

Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. (2003), 'The evolution of altruistic punishment', *Proceedings of the National Academy of Sciences* **100**(6), 3531–3535.

Boyd, R. & Richerson, P. J. (1988), 'The evolution of reciprocity in sizable groups', *Journal of Theoretical Biology* **132**(3), 337–356.

Boyd, R. & Richerson, P. J. (1992), 'Punishment allows the evolution of cooperation (or anything else) in sizable groups', *Ethology and Sociobiology* **13**(3), 171–195.

Boyd, R., Richerson, P. J. & Henrich, J. (2011), 'The cultural niche: Why social learning is essential for human adaptation', *Proceedings of the National Academy of Sciences* **108**(Supplement 2), 10918–10925.

Boyer, P. (2001), *Religion explained: The evolutionary origins of religious thought*, Basic Books.

Brewer, M. B. (1991), 'The social self: On being the same and different at the same time', *Personality and Social Psychology Bulletin* **17**(5), 475–482.

Broockman, D. E. & Butler, D. M. (2017), 'The causal effects of elite position-taking on voter attitudes: Field experiments with elite communication', *American Journal of Political Science* **61**(1), 208–221.

Brown, J. D. (2012), 'Understanding the better than average effect: Motives (still) matter', *Personality and Social Psychology Bulletin* **38**(2), 209–219.

Burnham, T. C. & Johnson, D. D. (2005), 'The biological and evolutionary logic of human cooperation', *Analyse & Kritik* **27**(1), 113–135.

Burton-Chellew, M. N., El Mouden, C. & West, S. A. (2016), 'Conditional cooperation and confusion in public-goods experiments', *Proceedings of the National Academy of Sciences* **113**(5), 1291–1296.

Burton-Chellew, M. N., Nax, H. H. & West, S. A. (2015), 'Payoff-based learning explains the decline in cooperation in public goods games', *Proceedings of the Royal Society B: Biological Sciences* **282**(1801), 20142678.

Burton-Chellew, M. N. & West, S. A. (2013), 'Prosocial preferences do not explain human cooperation in public-goods games', *Proceedings of the National Academy of Sciences* **110**(1), 216–221.

Burton-Chellew, M. N. & West, S. A. (2021), 'Payoff-based learning best explains the rate of decline in cooperation across 237 public-goods games', *Nature Human Behaviour* **5**(10), 1330–1338.

Burum, B., Nowak, M. A. & Hoffman, M. (2020), 'An evolutionary explanation for ineffective altruism', *Nature Human Behaviour* **4**(12), 1245–1257.

Camerer, C. F. (1997), 'Progress in behavioral game theory', *Journal of Economic Perspectives* **11**(4), 167–188.

Camerer, C. F. & Ho, T.-H. (2015), Behavioral game theory experiments and modeling, *in* '*Handbook of Game Theory with Economic Applications*', Vol. 4, Elsevier, pp. 517–573.

Carlo, G., Okun, M. A., Knight, G. P. & de Guzman, M. R. T. (2005), 'The interplay of traits and motives on volunteering: Agreeableness, extraversion and prosocial value motivation', *Personality and Individual Differences* **38**(6), 1293–1305.

Chambers, J. R. & Windschitl, P. D. (2004), 'Biases in social comparative judgments: the role of nonmotivated factors in above-average and comparative-optimism effects.', *Psychological Bulletin* **130**(5), 813.

Chambers, J. R., Windschitl, P. D. & Suls, J. (2003), 'Egocentrism, event frequency, and comparative optimism: When what happens frequently is "more likely to happen to me"', *Personality and Social Psychology Bulletin* **29**(11), 1343–1356.

Charness, G. & Chen, Y. (2020), 'Social identity, group behavior, and teams', *Annual Review of Economics* **12**, 691–713.

Charness, G. & Rabin, M. (2002), 'Understanding social preferences with simple tests', *The Quarterly Journal of Economics* **117**(3), 817–869.

Charness, G., Rustichini, A. & Van de Ven, J. (2018), 'Self-confidence and strategic behavior', *Experimental Economics* **21**(1), 72–98.

Chater, N. & Loewenstein, G. (2022), 'The i-frame and the s-frame: How focusing on the individual-level solutions has led behavioral public policy astray', *Available at SSRN 4046264* .

Chudek, M. & Henrich, J. (2011), 'Culture–gene coevolution, norm-psychology and the emergence of human prosociality', *Trends in Cognitive Sciences* **15**(5), 218–226.

Cisek, P. (2019), 'Resynthesizing behavior through phylogenetic refinement', *Attention, Perception, & Psychophysics* **81**(7), 2265–2287.

Cobb-Clark, D. A. & Tan, M. (2011), 'Noncognitive skills, occupational attainment, and relative wages', *Labour Economics* **18**(1), 1–13.

Connors, E. C. (2020), 'The social dimension of political values', *Political Behavior* **42**(3), 961–982.

Conte, A., Levati, M. V. & Montinari, N. (2019), 'Experience in public goods experiments', *Theory and Decision* **86**(1), 65–93.

Corriveau, K. & Harris, P. L. (2009), 'Choosing your informant: Weighing familiarity and recent accuracy', *Developmental Science* **12**(3), 426–437.

Coutts, A. (2019), 'Good news and bad news are still news: Experimental evidence on belief updating', *Experimental Economics* **22**, 369–395.

Coyne, J. C. & Tennen, H. (2010), 'Positive psychology in cancer care: Bad science, exaggerated claims, and unproven medicine', *Annals of Behavioral Medicine* **39**(1), 16–26.

Crawford, V. P. & Sobel, J. (1982), 'Strategic information transmission', *Econometrica: Journal of the Econometric Society* pp. 1431–1451.

Dana, J., Cain, D. M. & Dawes, R. M. (2006), 'What you don't know won't hurt me: Costly (but quiet) exit in dictator games', *Organizational Behavior and Human Decision Processes* **100**(2), 193–201.

Dana, J., Weber, R. A. & Kuang, J. X. (2007), 'Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness', *Economic Theory* **33**(1), 67–80.

Darwin, C. (1871), *The descent of man, and selection in relation to sex*, London: John Murray.

Dawes, C. T. & Weinschenk, A. C. (2020), 'On the genetic basis of political orientation', *Current Opinion in Behavioral Sciences* **34**, 173–178.

Dawkins, R. (1976), *The selfish gene*, Oxford University Press.

Dawkins, R. (1979), 'Twelve misunderstandings of kin selection', *Zeitschrift für Tierpsychologie* **51**(2), 184–200.

Dawkins, R. (2006), *The god delusion*, New York: Houghton Mifflin.

De Fruyt, F. & Mervielde, I. (1999), 'Riasec types and big five traits as predictors of employment status and nature of employment', *Personnel Psychology* **52**(3), 701–727.

DellaVigna, S. (2009), 'Psychology and economics: Evidence from the field', *Journal of Economic Literature* **47**(2), 315–72.

DellaVigna, S., List, J. A. & Malmendier, U. (2012), 'Testing for altruism and social pressure in charitable giving', *The Quarterly Journal of Economics* **127**(1), 1–56.

Denissen, J. J., Bleidorn, W., Hennecke, M., Luhmann, M., Orth, U., Specht, J. & Zimmermann, J. (2018), 'Uncovering the power of personality to shape income', *Psychological Science* **29**(1), 3–13.

Denissen, J. J., Penke, L., Schmitt, D. P. & Van Aken, M. A. (2008), 'Self-esteem reactions to social interactions: evidence for sociometer mechanisms across days, people, and nations.', *Journal of Personality and Social Psychology* **95**(1), 181.

DeScioli, P., Christner, J. & Kurzban, R. (2011), 'The omission strategy', *Psychological Science* **22**(4), 442–446.

DeScioli, P. & Karpoff, R. (2015), 'People's judgments about classic property law cases', *Human Nature* **26**(2), 184–209.

DeScioli, P. & Kurzban, R. (2013), 'A solution to the mysteries of morality.', *Psychological Bulletin* **139**(2), 477.

DeScioli, P. & Wilson, B. J. (2011), 'The territorial foundations of human property', *Evolution and Human Behavior* **32**(5), 297–304.

Di Tella, R., Perez-Truglia, R., Babino, A. & Sigman, M. (2015), 'Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism', *American Economic Review* **105**(11), 3416–42.

Dickinson, A. & Balleine, B. (2010), 'Hedonics: the cognitive-motivational interface', *Pleasures of the Brain* pp. 74–84.

Douenne, T. & Fabre, A. (2022), 'Yellow vests, pessimistic beliefs, and carbon tax aversion', *American Economic Journal: Economic Policy* **14**(1), 81–110.

Drobner, C. (2022), 'Motivated beliefs and anticipation of uncertainty resolution', *American Economic Review: Insights* **4**(1), 89–105.

Druckman, J. N., Peterson, E. & Slothuus, R. (2013), 'How elite partisan polarization affects public opinion formation', *American Political Science Review* **107**(1), 57–79.

Dunning, D., Meyerowitz, J. A. & Holzberg, A. D. (1989), 'Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability.', *Journal of Personality and Social Psychology* **57**(6), 1082.

Dziuda, W. & Salas, C. (2018), 'Communication with detectable deceit', *Available at SSRN 3234695* .

Ecklund, E. H., Johnson, D. R., Scheitle, C. P., Matthews, K. R. & Lewis, S. W. (2016), 'Religion among scientists in international context: A new study of scientists in eight regions', *Socius* **2**, 2378023116664353.

Ecklund, E. H. & Park, J. Z. (2009), 'Conflict between religion and science among academic scientists?', *Journal for the Scientific Study of Religion* **48**(2), 276–292.

Ecklund, E. H. & Scheitle, C. P. (2007), 'Religion among academic scientists: Distinctions, disciplines, and demographics', *Social Problems* **54**(2), 289–307.

Egan, P. J. (2020), 'Identity as dependent variable: How americans shift their identities to align with their politics', *American Journal of Political Science* **64**(3), 699–716.

Eil, D. & 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. & Dunning, D. (2000), 'Feeling" holier than thou": are self-serving assessments produced by errors in self-or social prediction?', *Journal of Personality and Social Psychology* **79**(6), 861.

Ertac, S. (2011), 'Does self-relevance affect information processing? experimental evidence on the response to performance and non-performance feedback', *Journal of Economic Behavior & Organization* **80**, 532–545.

Everitt, B. J. & Robbins, T. W. (2005), 'Neural systems of reinforcement for drug addiction: from actions to habits to compulsion', *Nature Neuroscience* **8**(11), 1481–1489.

Exley, C. L. (2016), 'Excusing selfishness in charitable giving: The role of risk', *The Review of Economic Studies* **83**(2), 587–628.

Exley, C. L. & Kessler, J. B. (2019), Motivated errors, Technical report, National Bureau of Economic Research.

Fehr, E. (2004), 'Don't lose your reputation', *Nature* **432**(7016), 449–450.

Fehr, E. & Fischbacher, U. (2002), 'Why social preferences matter–the impact of non-selfish motives on competition, cooperation and incentives', *The Economic Journal* **112**(478), C1–C33.

Fehr, E. & Fischbacher, U. (2003), 'The nature of human altruism', *Nature* **425**(6960), 785–791.

Fehr, E. & Gächter, S. (2002), 'Altruistic punishment in humans', *Nature* **415**(6868), 137–140.

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

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

Fleeson, W. & Wilt, J. (2010), 'The relevance of big five trait content in behavior to subjective authenticity: Do high levels of within-person behavioral variability undermine or enable authenticity achievement?', *Journal of Personality* **78**(4), 1353–1382.

Frey, U. & Voland, E. (2011), 'The evolutionary route to self-deception: Why offensive versus defensive strategy might be a false alternative', *Behavioral and Brain Sciences* **34**(1), 21.

Friedman, O. (2008), 'First possession: An assumption guiding inferences about who owns what', *Psychonomic Bulletin & Review* **15**(2), 290–295.

Friedman, O. & Neary, K. R. (2008), 'Determining who owns what: Do children infer ownership from first possession?', *Cognition* **107**(3), 829–849.

Fries, T., Gneezy, U., Kajackaite, A. & Parra, D. (2021), 'Observability and lying', *Journal of Economic Behavior & Organization* **189**, 132–149.

Frimer, J. A., Skitka, L. J. & Motyl, M. (2017), 'Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions', *Journal of Experimental Social Psychology* **72**, 1–12.

Fudenberg, D. (2006), 'Advancing beyond advances in behavioral economics', *Journal of Economic Literature* **44**(3), 694–711.

Fudenberg, D. & Levine, D. K. (1993), 'Self-confirming equilibrium', *Econometrica: Journal of the Econometric Society* pp. 523–545.

Fudenberg, D. & Levine, D. K. (1998), *The theory of learning in games*, MIT press.

Fudenberg, D. & Tirole, J. (1991), *Game theory*, MIT press.

Funk, C. L., Smith, K. B., Alford, J. R., Hibbing, M. V., Eaton, N. R., Krueger, R. F., Eaves, L. J. & Hibbing, J. R. (2013), 'Genetic and environmental transmission of political orientations', *Political Psychology* **34**(6), 805–819.

Funk, P. (2010), 'Social incentives and voter turnout: evidence from the swiss mail ballot system', *Journal of the European Economic Association* **8**(5), 1077–1103.

Gächter, S., Herrmann, B. & Thöni, C. (2010), 'Culture and cooperation', *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**(1553), 2651–2661.

Gächter, S. & Schulz, J. F. (2016), 'Intrinsic honesty and the prevalence of rule violations across societies', *Nature* **531**(7595), 496–499.

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

Gerber, A. S., Huber, G. A., Doherty, D., Dowling, C. M. & Ha, S. E. (2010), 'Personality and political attitudes: Relationships across issue domains and political contexts', *American Political Science Review* **104**(1), 111–133.

Gershman, S. J. (2019), 'How to never be wrong', *Psychonomic Bulletin & Review* **26**(1), 13–28.

Gervais, W. M., Najle, M. B. & Caluori, N. (2021), 'The origins of religious disbelief: A dual inheritance approach', *Social Psychological and Personality Science* **12**(7), 1369–1379.

Gintis, H. (2000), 'Strong reciprocity and human sociality', *Journal of Theoretical Biology* **206**(2), 169–179.

Gintis, H. (2007), 'The evolution of private property', *Journal of Economic Behavior & Organization* **64**(1), 1–16.

Gintis, H., Bowles, S., Boyd, R. & Fehr, E. (2003), 'Explaining altruistic behavior in humans', *Evolution and Human Behavior* **24**(3), 153–172.

Gneezy, U., Kajackaite, A. & Sobel, J. (2018), 'Lying aversion and the size of the lie', *American Economic Review* **108**(2), 419–53.

Goldstein, N. J., Cialdini, R. B. & Griskevicius, V. (2008), 'A room with a viewpoint: Using social norms to motivate environmental conservation in hotels', *Journal of Consumer Research* **35**(3), 472–482.

Golman, R. (2022), 'Acceptable discourse: Social norms of beliefs and opinions', *Available at SSRN 4160955* .

Golman, R., Hagmann, D. & Loewenstein, G. (2017), 'Information avoidance', *Journal of Economic Literature* **55**(1), 96–135.

Gould, E. D. & Klor, E. F. (2019), 'Party hacks and true believers: The effect of party affiliation on political preferences', *Journal of Comparative Economics* **47**(3), 504–524.

Greenberg, J., Gupta, S. & Luo, X. (2009), 'Mutually acceptable courses of action', *Economic Theory* **40**(1), 91–112.

Greenberg, J., Pyszczynski, T. & Solomon, S. (1986), The causes and consequences of a need for self-esteem: A terror management theory, *in* '*Public Self and Private Self*', Springer, pp. 189–212.

Grossman, Z. & Van der Weele, J. (2017), 'Self-image and willful ignorance in social decisions', *Journal of the European Economic Association* **15**(1), 173–217.

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

Güth, W., Schmittberger, R. & Schwarze, B. (1982), 'An experimental analysis of ultimatum bargaining', *Journal of Economic Behavior & Organization* **3**(4), 367–388.

Hamilton, W. (1964), 'The genetical evolution of social behaviour - i', *Journal of Theoretical Biology* **7**(1), 1–16.

Hart, S. & Tauman, Y. (2004), 'Market crashes without external shocks', *The Journal of Business* **77**(1), 1–8.

Hatemi, P. K., Funk, C. L., Medland, S. E., Maes, H. M., Silberg, J. L., Martin, N. G. & Eaves, L. J. (2009), 'Genetic and environmental transmission of political attitudes over a life time', *The Journal of Politics* **71**(3), 1141–1156.

Heine, S. J. (2001), 'Self as cultural product: An examination of east asian and north american selves', *Journal of Personality* **69**(6), 881–905.

Heine, S. J. & Hamamura, T. (2007), 'In search of east asian self-enhancement', *Personality and Social Psychology Review* **11**(1), 4–27.

Heine, S. J., Lehman, D. R., Markus, H. R. & Kitayama, S. (1999), 'Is there a universal need for positive self-regard?', *Psychological Review* **106**(4), 766.

Henrich, J. (2015), 'Culture and social behavior', *Current Opinion in Behavioral Sciences* **3**, 84–89.

Henrich, J. (2020), *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous*, Farrar, Straus and Giroux.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. & McElreath, R. (2001), 'In search of homo economicus: behavioral experiments in 15 small-scale societies', *American Economic Review* **91**(2), 73–78.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E. & Henrich, N. (2010), 'Markets, religion, community size, and the evolution of fairness and punishment', *Science* **327**(5972), 1480–1484.

Henrich, J. & Henrich, N. (2010), 'The evolution of cultural adaptations: Fijian food taboos protect against dangerous marine toxins', *Proceedings of the Royal Society B: Biological Sciences* **277**(1701), 3715–3724.

Herrmann, B., Thöni, C. & Gächter, S. (2008), 'Antisocial punishment across societies', *Science* **319**(5868), 1362–1367.

Hoff, K. & Pandey, P. (2006), 'Discrimination, social identity, and durable inequalities', *American Economic Review* **96**(2), 206–211.

Hoffman, M., Aygun Dalkiran, N., Sigstad, H. & Yoeli, E. (2018), 'Coordinated enforcement', *Working Paper, available at https://sites.google.com/site/hoffmanmoshe/* .

Hoffman, M. & Yoeli, E. (2022), *Hidden Games: The surprising power of game theory to explain irrational human behavior*, Basic Books.

Hoffman, M., Yoeli, E. & Navarrete, C. D. (2016), Game theory and morality, *in* '*The Evolution of Morality*', Springer, pp. 289–316.

Hopfensitz, A. & Reuben, E. (2009), 'The importance of emotions for the effectiveness of social punishment', *The Economic Journal* **119**(540), 1534–1559.

House, B. R., Kanngiesser, P., Barrett, H. C., Broesch, T., Cebioglu, S., Crittenden, A. N., Erut, A., Lew-Levy, S., Sebastian-Enesco, C., Smith, A. M., Yilmaz, S. & Silk, J. B. (2020), 'Universal norm psychology leads to societal diversity in prosocial behaviour and development', *Nature Human Behaviour* **4**(1), 36–44.

House, B. R., Silk, J. B., Henrich, J., Barrett, H. C., Scelza, B. A., Boyette, A. H., Hewlett, B. S., McElreath, R. & Laurence, S. (2013), 'Ontogeny of prosocial behavior across diverse societies', *Proceedings of the National Academy of Sciences* **110**(36), 14586–14591.

Huckfeldt, R., Mondak, J. J., Hayes, M., Pietryka, M. T. & Reilly, J. (2013), Networks, interdependence, and social influence in politics, *in* '*The Oxford Handbook of Political Psychology*', Oxford University Press.

Hufer, A., Kornadt, A. E., Kandler, C. & Riemann, R. (2020), 'Genetic and environmental variation in political orientation in adolescence and early adulthood: A nuclear twin family analysis.', *Journal of Personality and Social Psychology* **118**(4), 762.

Ickes, W., Snyder, M. & Garcia, S. (1997), Personality influences on the choice of situations, *in* '*Handbook of Personality Psychology*', Elsevier, pp. 165–195.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N. & Westwood, S. J. (2019), 'The origins and consequences of affective polarization in the united states', *Annual Review of Political Science* **22**, 129–146.

Jacoby, W. G. (2014), 'Is there a culture war? Conflicting value structures in American public opinion', *American Political Science Review* **108**(4), 754–771.

Johnson, D. (2004), *Overconfidence and war: The havoc and glory of positive illusions*, Harvard University Press.

Johnstone, R. A. & Grafen, A. (1993), 'Dishonesty and the handicap principle', *Animal Behaviour* **46**(4), 759–764.

Jordan, J. J., Hoffman, M., Bloom, P. & Rand, D. G. (2016), 'Third-party punishment as a costly signal of trustworthiness', *Nature* **530**(7591), 473–476.

Jordan, J. J. & Rand, D. G. (2017), 'Third-party punishment as a costly signal of high continuation probabilities in repeated games', *Journal of Theoretical Biology* **421**, 189–202.

Kahan, D. M. (2012), 'Ideology, motivated reasoning, and cognitive reflection: An experimental study', *Judgment and Decision Making* **8**, 407–24.

Kanngiesser, P., Gjersoe, N. & Hood, B. M. (2010), 'The effect of creative labor on property-ownership transfer by preschool children and adults', *Psychological Science* **21**(9), 1236–1241.

Karlan, D. & List, J. A. (2007), 'Does price matter in charitable giving? evidence from a large-scale natural field experiment', *American Economic Review* **97**(5), 1774–1793.

Kartik, N. (2009), 'Strategic communication with lying costs', *The Review of Economic Studies* **76**(4), 1359–1395.

Kay, T., Keller, L. & Lehmann, L. (2020), 'The evolution of altruism and the serial rediscovery of the role of relatedness', *Proceedings of the National Academy of Sciences* **117**(46), 28894–28898.

Keizer, K., Lindenberg, S. & Steg, L. (2008), 'The spreading of disorder', *Science* **322**(5908), 1681–1685.

Kennedy, J. A., Anderson, C. & Moore, D. A. (2013), 'When overconfidence is revealed to others: Testing the status-enhancement theory of overconfidence', *Organizational Behavior and Human Decision Processes* **122**(2), 266–279.

Khalmetski, K. & Sliwka, D. (2019), 'Disguising lies—image concerns and partial lying in cheating games', *American Economic Journal: Microeconomics* **11**(4), 79–110.

Kirchsteiger, G. (1994), 'The role of envy in ultimatum games', *Journal of Economic Behavior & Organization* **25**(3), 373–389.

Klar, S. (2014), 'Partisanship in a social setting', *American Journal of Political Science* **58**(3), 687–704.

Kobayashi, C. & Brown, J. D. (2003), 'Self-esteem and self-enhancement in japan and america', *Journal of Cross-Cultural Psychology* **34**(5), 567–580.

Kokko, H., Johnstone, R. & Clutton-Brock, T. (2001), 'The evolution of cooperative breeding through group augmentation', *Proceedings of the Royal Society of London. Series B: Biological Sciences* **268**(1463), 187–196.

Kokko, H., López-Sepulcre, A. & Morrell, L. J. (2006), 'From hawks and doves to self-consistent games of territorial behavior', *The American Naturalist* **167**(6), 901–912.

Kool, W. & Botvinick, M. (2018), 'Mental labour', *Nature Human Behaviour* **2**(12), 899–908.

Kool, W., Cushman, F. A. & Gershman, S. J. (2018), 'Competition and cooperation between multiple reinforcement learning systems', *Goal-Directed Decision Making* pp. 153–178.

Kool, W., Gershman, S. J. & Cushman, F. A. (2017), 'Cost-benefit arbitration between multiple reinforcement-learning systems', *Psychological Science* **28**(9), 1321–1333.

Kool, W., McGuire, J. T., Rosen, Z. B. & Botvinick, M. M. (2010), 'Decision making and the avoidance of cognitive demand.', *Journal of Experimental Psychology: General* **139**(4), 665.

Köszegi, B. (2006), 'Ego utility, overconfidence, and task choice', *Journal of the European Economic Association* **4**(4), 673–707.

Kraft-Todd, G., Yoeli, E., Bhanot, S. & Rand, D. (2015), 'Promoting cooperation in the field', *Current Opinion in Behavioral Sciences* **3**, 96–101.

Krakauer, A. H. (2005), 'Kin selection and cooperative courtship in wild turkeys', *Nature* **434**(7029), 69–72.

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

Kruger, J. (1999), 'Lake wobegon be gone! the" below-average effect" and the egocentric nature of comparative ability judgments.', *Journal of Personality and Social Psychology* **77**(2), 221.

Kümmerli, R., Burton-Chellew, M. N., Ross-Gillespie, A. & West, S. A. (2010), 'Resistance to extreme strategies, rather than prosocial preferences, can explain human cooperation in public goods games', *Proceedings of the National Academy of Sciences* **107**(22), 10125–10130.

Kurland, J. A. & Gaulin, S. J. (2005), 'Cooperation and conflict among kin', *The Handbook of Evolutionary Psychology, ed. DM Buss* pp. 447–82.

Kurzban, R. (2012), *Why everyone (else) is a hypocrite: Evolution and the modular mind*, Princeton University Press.

Kurzban, R. & Aktipis, A. (2007), 'Modularity and the social mind: Are psychologists too self-ish?', *Personality and Social Psychology Review* **11**(2), 131–149.

Kurzban, R., Burton-Chellew, M. N. & West, S. A. (2015), 'The evolution of altruism in humans', *Annual Review of Psychology* **66**, 575–599.

Lachmann, M., Szamado, S. & Bergstrom, C. T. (2001), 'Cost and conflict in animal signals and human language', *Proceedings of the National Academy of Sciences* **98**(23), 13189–13194.

Laibson, D. & List, J. A. (2015), 'Principles of (behavioral) economics', *American Economic Review* **105**(5), 385–90.

Lamba, S. & Mace, R. (2011), 'Demography and ecology drive variation in cooperation across human populations', *Proceedings of the National Academy of Sciences* **108**(35), 14426–14430.

Lamba, S. & Nityananda, V. (2014), 'Self-deceived individuals are better at deceiving others', *PloS One* **9**(8), e104562.

Lanman, J. A. (2012), 'The importance of religious displays for belief acquisition and secularization', *Journal of Contemporary Religion* **27**(1), 49–65.

Lazer, D., Rubineau, B., Chetkovich, C., Katz, N. & Neblo, M. (2010), 'The coevolution of networks and political attitudes', *Political Communication* **27**(3), 248–274.

Leary, M. R. (2004), 'What is the self? a plea for clarity', *Self and Identity* **3**(1), 1–3.

Leary, M. R., Tambor, E. S., Terdal, S. K. & Downs, D. L. (1995), 'Self-esteem as an interpersonal monitor: The sociometer hypothesis.', *Journal of Personality and Social Psychology* **68**(3), 518.

LeDoux, J. (2012), 'Rethinking the emotional brain', *Neuron* **73**(4), 653–676.

LeDoux, J. E. (2015), *Anxious: Using the brain to understand and treat fear and anxiety*, Penguin.

LeDoux, J. E. & Brown, R. (2017), 'A higher-order theory of emotional consciousness', *Proceedings of the National Academy of Sciences* **114**(10), E2016–E2025.

Lehtonen, J. (2016), 'Multilevel selection in kin selection language', *Trends in Ecology & Evolution* **31**(10), 752–762.

Levine, D. K. (1998), 'Modeling altruism and spitefulness in experiments', *Review of Economic Dynamics* **1**(3), 593–622.

Levitan, L. C. & Verhulst, B. (2016), 'Conformity in groups: The effects of others' views on expressed attitudes and attitude change', *Political Behavior* **38**(2), 277–315.

Lieberman, D. & Lobel, T. (2012), 'Kinship on the kibbutz: Coresidence duration predicts altruism, personal sexual aversions and moral attitudes among communally reared peers', *Evolution and Human Behavior* **33**(1), 26–34.

Lieberman, D., Tooby, J. & Cosmides, L. (2007), 'The architecture of human kin detection', *Nature* **445**(7129), 727–731.

Loewenstein, G. (1987), 'Anticipation and the valuation of delayed consumption', *The Economic Journal* **97**(387), 666–684.

Loewenstein, G. & Molnar, A. (2018), 'The renaissance of belief-based utility in economics', *Nature Human Behaviour* **2**(3), 166–167.

Logg, J. M., Haran, U. & Moore, D. A. (2018), 'Is overconfidence a motivated bias? experimental evidence.', *Journal of Experimental Psychology: General* **147**(10), 1445.

Loury, G. C. (1994), 'Self-censorship in public discourse: A theory of "political correctness" and related phenomena', *Rationality and Society* **6**(4), 428–461.

Lowes, S., Nunn, N., Robinson, J. A. & Weigel, J. L. (2017), 'The evolution of culture and institutions: Evidence from the kuba kingdom', *Econometrica* **85**(4), 1065–1091.

Malmendier, U. & Tate, G. (2005), 'Ceo overconfidence and corporate investment', *The Journal of Finance* **60**(6), 2661–2700.

Marshall, J. A. (2011), 'Group selection and kin selection: formally equivalent approaches', *Trends in Ecology & Evolution* **26**(7), 325–332.

Marshall, J. A., Trimmer, P. C., Houston, A. I. & McNamara, J. M. (2013), 'On evolutionary explanations of cognitive biases', *Trends in Ecology & Evolution* **28**(8), 469–473.

Martin, G. J. & Webster, S. W. (2020), 'Does residential sorting explain geographic polarization?', *Political Science Research and Methods* **8**(2), 215–231.

Maynard Smith, J. (1982), *Evolution and the Theory of Games*, Cambridge university press.

Mayrl, D. & Uecker, J. E. (2011), 'Higher education and religious liberalization among young adults', *Social Forces* **90**(1), 181–208.

McAuliffe, W. H., Burton-Chellew, M. N. & McCullough, M. E. (2019), 'Cooperation and learning in unfamiliar situations', *Current Directions in Psychological Science* **28**(5), 436–440.

McAuliffe, W. H., Forster, D. E., Pedersen, E. J. & McCullough, M. E. (2018), 'Experience with anonymous interactions reduces intuitive cooperation', *Nature Human Behaviour* **2**(12), 909–914.

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

McGuigan, N., Makinson, J. & Whiten, A. (2011), 'From over-imitation to super-copying: Adults imitate causally irrelevant aspects of tool use with higher fidelity than young children', *British Journal of Psychology* **102**(1), 1–18.

McKay, R. T. & Dennett, D. C. (2009), 'The evolution of misbelief', *Behavioral & Brain Sciences* **32**(6), 493–510.

Melnikoff, D. E. & Strohminger, N. (2020), 'The automatic influence of advocacy on lawyers and novices', *Nature Human Behaviour* pp. 1–7.

Messick, D. M., Bloom, S., Boldizar, J. P. & Samuelson, C. D. (1985), 'Why we are fairer than others', *Journal of Experimental Social Psychology* **21**(5), 480–500.

Minsky, M. (1988), *Society of mind*, Simon and Schuster.

Mitchell, K. J. (2020), *Innate: How the wiring of our brains shapes who we are*, Princeton University Press.

Möbius, M. M., Niederle, M., Niehaus, P. & Rosenblat, T. S. (2022), 'Managing self-confidence: Theory and experimental evidence', *Management Science* .

Molleman, L., Kölle, F., Starmer, C. & Gächter, S. (2019), 'People prefer coordinated punishment in cooperative interactions', *Nature Human Behaviour* **3**(11), 1145–1153.

Molnar, A. & Loewenstein, G. (2021), 'Thoughts and players: An introduction to old and new economic perspectives on beliefs', *The Science of Beliefs: A multidisciplinary Approach. Cambridge University Press. Edited by Julien Musolino, Joseph Sommer, and Pernille Hemmer* .

Moore, D. A. & Healy, P. J. (2008), 'The trouble with overconfidence.', *Psychological Review* **115**(2), 502.

Morgan, T. J., Rendell, L. E., Ehn, M., Hoppitt, W. & Laland, K. N. (2012), 'The evolutionary basis of human social learning', *Proceedings of the Royal Society B: Biological Sciences* **279**(1729), 653–662.

Mutz, D. C. & Mondak, J. J. (2006), 'The workplace as a context for cross-cutting political discourse', *The Journal of Politics* **68**(1), 140–155.

Nettle, D. (2015), *Tyneside neighbourhoods: Deprivation, social life and social behaviour in one British city*, Open Book Publishers.

Nettle, D., Colléony, A. & Cockerill, M. (2011), 'Variation in cooperative behaviour within a single city', *PloS One* **6**(10).

Nettle, D., Gibson, M. A., Lawson, D. W. & Sear, R. (2013), 'Human behavioral ecology: current research and future prospects', *Behavioral Ecology* **24**(5), 1031–1040.

Nettle, D., Pepper, G. V., Jobling, R. & Schroeder, K. B. (2014), 'Being there: a brief visit to a neighbourhood induces the social attitudes of that neighbourhood', *PeerJ* **2**, e236.

Nisbett, R. E., Peng, K., Choi, I. & Norenzayan, A. (2001), 'Culture and systems of thought: holistic versus analytic cognition.', *Psychological Review* **108**(2), 291.

Norenzayan, A. (2013), *Big gods: How religion transformed cooperation and conflict*, Princeton University Press.

Norenzayan, A., Shariff, A. F., Gervais, W. M., Willard, A. K., McNamara, R. A., Slingerland, E. & Henrich, J. (2016), 'The cultural evolution of prosocial religions', *Behavioral & Brain Sciences* **39**.

Nowak, M. A. & Sigmund, K. (1998), 'Evolution of indirect reciprocity by image scoring', *Nature* **393**(6685), 573–577.

Null, C. (2011), 'Warm glow, information, and inefficient charitable giving', *Journal of Public Economics* **95**(5-6), 455–465.

Nyhan, B., Porter, E., Reifler, J. & Wood, T. J. (2019), 'Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability', *Political Behavior* pp. 1–22.

Nyhan, B. & Reifler, J. (2010), 'When corrections fail: The persistence of political misperceptions', *Political Behavior* **32**(2), 303–330.

Panchanathan, K. & Boyd, R. (2004), 'Indirect reciprocity can stabilize cooperation without the second-order free rider problem', *Nature* **432**(7016), 499–502.

Pedregon, C. A., Farley, R. L., Davis, A., Wood, J. M. & Clark, R. D. (2012), 'Social desirability, personality questionnaires, and the "better than average" effect', *Personality and Individual Differences* **52**(2), 213–217.

Peysakhovich, A. & Rand, D. G. (2016), 'Habits of virtue: Creating norms of cooperation and defection in the laboratory', *Management Science* **62**(3), 631–647.

Pietraszewski, D., Curry, O. S., Petersen, M. B., Cosmides, L. & Tooby, J. (2015), 'Constituents of political cognition: Race, party politics, and the alliance detection system', *Cognition* **140**, 24–39.

Powers, S. T., van Schaik, C. P. & Lehmann, L. (2016), 'How institutions shaped the last major evolutionary transition to large-scale human societies', *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**(1687), 20150098.

Price, G. R. (1970), 'Selection and covariance.', *Nature* **227**, 520–521.

Price, P. C. & Stone, E. R. (2004), 'Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic', *Journal of Behavioral Decision Making* **17**(1), 39–57.

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

Rabin, M. (2013), 'An approach to incorporating psychology into economics', *American Economic Review* **103**(3), 617–22.

Raihani, N. J. & Bshary, R. (2015), 'Why humans might help strangers', *Frontiers in Behavioral Neuroscience* **9**, 39.

Rakoczy, H. & Schmidt, M. F. (2013), 'The early ontogeny of social norms', *Child Development Perspectives* **7**(1), 17–21.

Rand, D. G., Greene, J. D. & Nowak, M. A. (2012), 'Spontaneous giving and calculated greed', *Nature* **489**(7416), 427–430.

Rand, D. G. & Kraft-Todd, G. T. (2014), 'Reflection does not undermine self-interested prosociality', *Frontiers in Behavioral Neuroscience* **8**, 300.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A. & Greene, J. D. (2014), 'Social heuristics shape intuitive cooperation', *Nature Communications* **5**(1), 1–12.

Rand, D. G., Yoeli, E. & Hoffman, M. (2014), 'Harnessing reciprocity to promote cooperation and the provisioning of public goods', *Policy Insights from the Behavioral and Brain Sciences* **1**(1), 263–269.

Rich, P. & Zollman, K. J. (2016), 'Honesty through repeated interactions', *Journal of Theoretical Biology* **395**, 238–244.

Richerson, P., Baldini, R., Bell, A. V., Demps, K., Frost, K., Hillis, V., Mathew, S., Newton, E. K., Naar, N., Newson, L., Ross, C., Smaldino, P., Waring, T. & Zefferman, M. (2016), 'Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence', *Behavioral and Brain Sciences* **39**.

Rogers, T., Ternovski, J. & Yoeli, E. (2016), 'Potential follow-up increases private contributions to public goods', *Proceedings of the National Academy of Sciences* **113**(19), 5218–5220.

Rolls, E. (1999), *The Brain and Emotion*, New York: Oxford University Press.

Sah, S., Moore, D. A. & MacCoun, R. J. (2013), 'Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness', *Organizational Behavior and Human Decision Processes* **121**(2), 246–255.

Santos-Pinto, L. & Sobel, J. (2005), 'A model of positive self-image in subjective assessments', *American Economic Review* **95**(5), 1386–1402.

Savage, L. J. (1954), *The foundations of statistics.*, John Wiley & Sons.

Schaffner, B. F. & Roche, C. (2016), 'Misinformation and motivated reasoning: Responses to economic news in a politicized environment', *Public Opinion Quarterly* **81**(1), 86–110.

Schmidt, M. F. & Tomasello, M. (2012), 'Young children enforce social norms', *Current Directions in Psychological Science* **21**(4), 232–236.

Schroeder, K. B., Pepper, G. V. & Nettle, D. (2014), 'Local norms of cheating and the cultural evolution of crime and punishment: a study of two urban neighborhoods', *PeerJ* **2**, e450.

Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P. & Henrich, J. (2019), 'The church, intensive kinship, and global psychological variation', *Science* **366**(6466), eaau5141.

Schwardmann, P., Tripodi, E. & Van der Weele, J. J. (2022), 'Self-persuasion: Evidence from field experiments at international debating competitions', *American Economic Review* **112**(4), 1118–46.

Schwardmann, P. & Van der Weele, J. (2019), 'Deception and self-deception', *Nature Human Behaviour* **3**(10), 1055–1061.

Schwartz, D., Keenan, E. A., Imas, A. & Gneezy, A. (2019), 'Opting-in to prosocial incentives', *Organizational Behavior and Human Decision Processes* .

Scott-Phillips, T. C., Dickins, T. E. & West, S. A. (2011), 'Evolutionary theory and the ultimate–proximate distinction in the human behavioral sciences', *Perspectives on Psychological Science* **6**(1), 38–47.

Sedikides, C. (2022), 'Self-enhancement and physical health: A meta-analysis', *British Journal of Social Psychology* **00**, 1–17.

Sedikides, C., Gaertner, L. & Toguchi, Y. (2003), 'Pancultural self-enhancement.', *Journal of Personality and Social Psychology* **84**(1), 60.

Sedikides, C., Gaertner, L. & Vevea, J. L. (2005), 'Pancultural self-enhancement reloaded: A meta-analytic reply to heine (2005)', *Journal of Personality and Social Psychology* **89**(4), 539–551.

Sedikides, C., Gaertner, L. & Vevea, J. L. (2007), 'Inclusion of theory-relevant moderators yield the same conclusions as sedikides, gaertner, and vevea (2005): A meta-analytical reply to heine, kitayama, and hamamura (2007)', *Asian Journal of Social Psychology* **10**(2), 59–67.

Sedikides, C. & Gregg, A. P. (2008), 'Self-enhancement: Food for thought', *Perspectives on Psychological Science* **3**(2), 102–116.

Sedikides, C., Herbst, K. C., Hardin, D. P. & Dardis, G. J. (2002), 'Accountability as a deterrent to self-enhancement: The search for mechanisms', *Journal of Personality and Social Psychology* **83**(3), 592–605.

Segerstrom, S. C., Taylor, S. E., Kemeny, M. E. & Fahey, J. L. (1998), 'Optimism is associated with mood, coping, and immune change in response to stress.', *Journal of Personality and Social Psychology* **74**(6), 1646.

Sell, A., Sznycer, D., Al-Shawaf, L., Lim, J., Krauss, A., Feldman, A., Rascanu, R., Sugiyama, L., Cosmides, L. & Tooby, J. (2017), 'The grammar of anger: Mapping the computational architecture of a recalibrational emotion', *Cognition* **168**, 110–128.

Seth, A. (2021), *Being you: A new science of consciousness*, Penguin.

Shalvi, S., Dana, J., Handgraaf, M. J. & De Dreu, C. K. (2011), 'Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior', *Organizational Behavior and Human Decision Processes* **115**(2), 181–190.

Shayo, M. (2009), 'A model of social identity with an application to political economy: Nation, class, and redistribution', *American Political Science Review* **103**(2), 147–174.

Shayo, M. (2020), 'Social identity and economic policy', *Annual Review of Economics* .

Sheldon, K. M., Ryan, R. M., Rawsthorne, L. J. & Ilardi, B. (1997), 'Trait self and true self: Cross-role variation in the big-five personality traits and its relations with psychological authenticity and subjective well-being.', *Journal of Personality and Social Psychology* **73**(6), 1380.

Sherman, P. W. (1977), 'Nepotism and the evolution of alarm calls', *Science* **197**(4310), 1246–1253.

Shipman, A. S. & Mumford, M. D. (2011), 'When confidence is detrimental: Influence of overconfidence on leadership effectiveness', *The Leadership Quarterly* **22**(4), 649–665.

Silk, J. B., Kaldor, E. & Boyd, R. (2000), 'Cheap talk when interests conflict', *Animal Behaviour* **59**(2), 423–432.

Sinclair, B. (2012), *The social citizen: Peer networks and political behavior*, University of Chicago Press.

Singh, M. (2020), 'Subjective selection and the evolution of complex culture'.

Singh, M., Wrangham, R. & Glowacki, L. (2017), 'Self-interest and the design of rules', *Human Nature* **28**(4), 457–480.

Slothuus, R. & Bisgaard, M. (2021), 'How political parties shape public opinion in the real world', *American Journal of Political Science* **65**(4), 896–911.

Smaldino, P. E. (2019), 'Social identity and cooperation in cultural evolution', *Behavioural Processes* **161**, 108–116.

Smaldino, P. E., Lukaszewski, A., von Rueden, C. & Gurven, M. (2019), 'Niche diversity can explain cross-cultural differences in personality structure', *Nature Human Behaviour* **3**(12), 1276–1283.

Smith, A., Pedersen, E. J., Forster, D. E., McCullough, M. E. & Lieberman, D. (2017), 'Cooperation: The roles of interpersonal value and gratitude', *Evolution and Human Behavior* **38**(6), 695–703.

Smith, D. (2020), 'Cultural group selection and human cooperation: a conceptual and empirical review', *Evolutionary Human Sciences* **2**, 1–29.

Smith, J. M. (1964), 'Group selection and kin selection', *Nature* **201**(4924), 1145–1147.

Sng, O., Neuberg, S. L., Varnum, M. E. & Kenrick, D. T. (2018), 'The behavioral ecology of cultural psychological variation.', *Psychological Review* **125**(5), 714.

Solda, A., Ke, C., Page, L. & Von Hippel, W. (2019), 'Strategically delusional', *Experimental Economics* pp. 1–28.

Solms, M. (2021), *The hidden spring: A journey to the source of consciousness*, WW Norton & Company.

Spranca, M., Minsk, E. & Baron, J. (1991), 'Omission and commission in judgment and choice', *Journal of Experimental Social Psychology* **27**(1), 76–105.

Stagnaro, M. N., Arechar, A. A. & Rand, D. G. (2017), 'From good institutions to generous citizens: Top-down incentives to cooperate promote subsequent prosociality but not norm enforcement', *Cognition* **167**, 212–254.

Sterling, P. & Laughlin, S. (2015), *Principles of neural design*, MIT press.

Stigler, G. J. (1961), 'The economics of information', *Journal of Political Economy* **69**(3), 213–225.

Sutton, R. S. & Barto, A. G. (2018), *Reinforcement learning: An introduction*, MIT press.

Sweeny, K. & Krizan, Z. (2013), 'Sobering up: A quantitative review of temporal declines in expectations.', *Psychological Bulletin* **139**(3), 702.

Swire-Thompson, B., Ecker, U. K., Lewandowsky, S. & Berinsky, A. J. (2020), 'They might be a liar but they're my liar: Source evaluation and the prevalence of misinformation', *Political Psychology* **41**(1), 21–34.

Sznycer, D., Al-Shawaf, L., Bereby-Meyer, Y., Curry, O. S., De Smet, D., Ermer, E., Kim, S., Kim, S., Li, N. P. & Seal, M. F. L. (2017), 'Cross-cultural regularities in the cognitive architecture of pride', *Proceedings of the National Academy of Sciences* **114**(8), 1874–1879.

Sznycer, D., De Smet, D., Billingsley, J. & Lieberman, D. (2016), 'Coresidence duration and cues of maternal investment regulate sibling altruism across cultures.', *Journal of Personality and Social Psychology* **111**(2), 159.

Sznycer, D., Delton, A. W., Robertson, T. E., Cosmides, L. & Tooby, J. (2019), 'The ecological rationality of helping others: Potential helpers integrate cues of recipients' need and willingness to sacrifice', *Evolution and Human Behavior* **40**(1), 34–45.

Sznycer, D. & Lukaszewski, A. W. (2019), 'The emotion–valuation constellation: Multiple emotions are governed by a common grammar of social valuation', *Evolution and Human Behavior* **40**(4), 395–404.

Sznycer, D., Tooby, J., Cosmides, L., Porat, R., Shalvi, S. & Halperin, E. (2016), 'Shame closely tracks the threat of devaluation by others, even across cultures', *Proceedings of the National Academy of Sciences* **113**(10), 2625–2630.

Sznycer, D., Xygalatas, D., Agey, E., Alami, S., An, X.-F., Ananyeva, K. I., Atkinson, Q. D., Broitman, B. R., Conte, T. J. & Flores, C. (2018), 'Cross-cultural invariances in the architecture of shame', *Proceedings of the National Academy of Sciences* **115**(39), 9702–9707.

Tappin, B. M. & McKay, R. T. (2017), 'The illusion of moral superiority', *Social Psychological and Personality Science* **8**(6), 623–631.

Tappin, B. M., Pennycook, G. & Rand, D. G. (2020), 'Thinking clearly about causal inferences of politically motivated reasoning: Why paradigmatic study designs often undermine causal inference', *Current Opinion in Behavioral Sciences* **34**, 81–87.

Taylor, K. M. & Shepperd, J. A. (1998), 'Bracing for the worst: Severity, testing, and feedback timing as moderators of the optimistic bias', *Personality and Social Psychology Bulletin* **24**(9), 915–926.

Taylor, S. E. & Brown, J. D. (1988), 'Illusion and well-being: A social psychological perspective on mental health.', *Psychological Bulletin* **103**(2), 193.

Taylor, S. E., Kemeny, M. E., Reed, G. M., Bower, J. E. & Gruenewald, T. L. (2000), 'Psychological resources, positive illusions, and health.', *American Psychologist* **55**(1), 99.

Taylor, S. E., Lerner, J. S., Sherman, D. K., Sage, R. M. & McDowell, N. K. (2003), 'Are self-enhancing cognitions associated with healthy or unhealthy biological profiles?', *Journal of Personality and Social Psychology* **85**(4), 605.

Tenney, E. R., MacCoun, R. J., Spellman, B. A. & Hastie, R. (2007), 'Calibration trumps confidence as a basis for witness credibility', *Psychological Science* **18**(1), 46–50.

Tenney, E. R., Meikle, N. L., Hunsaker, D., Moore, D. A. & Anderson, C. (2019), 'Is overconfidence a social liability? the effect of verbal versus nonverbal expressions of confidence.', *Journal of Personality and Social Psychology* **116**(3), 396.

Tenney, E. R., Spellman, B. A. & MacCoun, R. J. (2008), 'The benefits of knowing what you know (and what you don't): How calibration affects credibility', *Journal of Experimental Social Psychology* **44**(5), 1368–1375.

Thorndike, E. L. (1927), 'The law of effect', *The American Journal of Psychology* **39**(1/4), 212–222.

Tice, D. M., Butler, J. L., Muraven, M. B. & Stillwell, A. M. (1995), 'When modesty prevails: Differential favorability of self-presentation to friends and strangers.', *Journal of Personality and Social Psychology* **69**(6), 1120.

Toelch, U., Bruce, M. J., Newson, L., Richerson, P. J. & Reader, S. M. (2014), 'Individual consistency and flexibility in human social information use', *Proceedings of the Royal Society B: Biological Sciences* **281**(1776), 20132864.

Toff, B. & Suhay, E. (2019), 'Partisan conformity, social identity, and the formation of policy preferences', *International Journal of Public Opinion Research* **31**(2), 349–367.

Tooby, J., Cosmides, L., Sell, A., Lieberman, D. & Sznycer, D. (2008), 'Internal regulatory variables and the design of human motivation: A computational and evolutionary approach', *Handbook of Approach and Avoidance Motivation* **15**, 251.

Trivers, R. (1971), 'The evolution of reciprocal altruism', *The Quarterly Review of Biology* **46**(1), 35–57.

Umbhauer, G. & Wolff, A. (2019), 'Individually-consistent sequential equilibrium', *BETA Working Paper N° 2019-39* .

Van Bavel, J. J. & Pereira, A. (2018), 'The partisan brain: An identity-based model of political belief', *Trends in Cognitive Sciences* **22**(3), 213–224.

Van den Steen, E. (2004), 'Rational overoptimism (and other biases)', *American Economic Review* **94**(4), 1141–1151.

Van Lange, P. A. & Sedikides, C. (1998), 'Being more honest but not necessarily more intelligent than others: Generality and explanations for the muhammad ali effect', *European Journal of Social Psychology* **28**(4), 675–680.

Visser, P. S. & Mirabile, R. R. (2004), 'Attitudes in the social context: The impact of social network composition on individual-level attitude strength.', *Journal of Personality and Social Psychology* **87**(6), 779.

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

Vullioud, C., Clément, F., Scott-Phillips, T. & Mercier, H. (2017), 'Confidence as an expression of commitment: Why misplaced expressions of confidence backfire', *Evolution and Human Behavior* **38**(1), 9–17.

Webster, M. S., Ligon, R. A. & Leighton, G. M. (2018), 'Social costs are an under-appreciated force for honest signalling in animal aggregations', *Animal Behaviour* **143**, 167–176.

Weinstein, N. D. (1980), 'Unrealistic optimism about future life events.', *Journal of Personality and Social Psychology* **39**(5), 806.

Wells, R., Ham, R. & Junankar, P. N. (2016), 'An examination of personality in occupational outcomes: antagonistic managers, careless workers and extraverted salespeople', *Applied Economics* **48**(7), 636–651.

West, S. A., El Mouden, C. & Gardner, A. (2011), 'Sixteen common misconceptions about the evolution of cooperation in humans', *Evolution and Human Behavior* **32**(4), 231–262.

West, S. A. & Gardner, A. (2010), 'Altruism, spite, and greenbeards', *Science* **327**(5971), 1341–1344.

West, S. A. & Gardner, A. (2013), 'Adaptation and inclusive fitness', *Current Biology* **23**(13), R577–R584.

West, S. A., Griffin, A. S. & Gardner, A. (2007), 'Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection', *Journal of Evolutionary Biology* **20**(2), 415–432.

Westbrook, A., Kester, D. & Braver, T. S. (2013), 'What is the subjective cost of cognitive effort? load, trait, and aging effects revealed by economic preference', *PloS One* **8**(7), e68210.

Wiessner, P. (2005), 'Norm enforcement among the ju/'hoansi bushmen', *Human Nature* **16**(2), 115–145.

Willard, A. K. & Cingl, L. (2017), 'Testing theories of secularization and religious belief in the czech republic and slovakia', *Evolution and Human Behavior* **38**(5), 604–615.

Williams, D. (2020), 'Socially adaptive belief', *Mind & Language* pp. 1–22.

Yoeli, E., Hoffman, M., Rand, D. G. & Nowak, M. A. (2013), 'Powering up with indirect reciprocity in a large-scale field experiment', *Proceedings of the National Academy of Sciences* **110**, 10424–10429.

Zell, E., Strickhouser, J. E., Sedikides, C. & Alicke, M. D. (2020), 'The better-than-average effect in comparative self-evaluation: A comprehensive review and meta-analysis.', *Psychological Bulletin* **146**(2), 118.

Ziano, I., Mok, P. Y. & Feldman, G. (2021), 'Replication and extension of alicke (1985) better-than-average effect for desirable and controllable traits', *Social Psychological and Personality Science* **12**(6), 1005–1017.