An Essay on the Foundations of Game Theory with Applications to the Theory of Public Goods

Inauguraldissertation zur Erlangung des akademischen Grades eines Doktors der Wirtschaftswissenschaften der Universität Mannheim

vorgelegt von

Justin Leduc

Herbstsemester 2019

Abteilungssprecher: Prof. Dr. Hans Peter Grüner

Referent: Prof. Dr. Thomas Tröger

Korreferent: Prof. Dr. Henrik Orzen

Tag der mündlichen Prüfung: 23. September 2019

À mes parents, Yasmine et Bernard.

Avant d'être la sœur du rêve, l'action est la fille de la rigueur.

Georges Canguilhem, Vie et mort de Jean Cavaillès.[1]

[1] "Action is the sister of dreams, but it is rigor's daughter first." Georges Canguilhem, Life and Death of Jean Cavaillès.

Contents

Acknowledgments

Idiosyncratic interests and obstinacy make life hard for supervisors and supervisees. Nonetheless, I am thankful to my supervisor, Thomas Tröger, who found the patience to stand my long—sometimes questionable—excursions. Thomas' illustrations of the remarkable reach of mainstream modelling operated as lighthouse signals: they reminded me not to loose sight of the shore. It is quite likely that, without them, I would have lost sight of it. I am also very grateful to Robert (Bob) Sugden, from the University of East Anglia. I feel really lucky that I could spend the month of April 2018 in Norwich and make Bob's acquaintance. His engaging discussions as well as his positive outlook on my research were real sources of motivation. They brought me an authoritative ground for believing in the prospects of this dissertation. Any person who has ever engaged in a long and difficult endeavor will know how valuable that is.

Further academic thanks go to Helge Rückert, Christophe Levêque, Henrik Orzen, and Carole Haritchabalet. Looking for feedback from a person with actual expertise in philosophy, I contacted Helge in February 2018. He showed great kindness by accepting to have a look at several obscure drafts of mine and did so quite helpfully, for he was keen on developing constructive critiques of my work. Christophe, my good friend and fellow social scientist, also took time to carefully read early drafts of Chapters 1 and 2 presented here. His support and honest feedback throughout the years were precious to me. I am very grateful to Henrik too, for accepting to take the role of second referee in my dissertation committee. And, lastly, I want to mention an older debt of gratitude I owe to Carole. If it weren't for her, I might not have considered international academia as an available alternative, I might not have had a chance to take my masters in Toulouse, and, at any rate, I would not have found the necessary self-confidence to undertake any of the two endeavors.

Three years into the program, I took a tough decision to delve into philosophy, a field in which I am not learned. I am happy with this decision, because I firmly believe that interesting work can be done at the frontier of microeconomics and philosophy. Come what may, the choice involved a significant lengthening of my studies. So, if it weren't for

the generous material support from the University of Mannheim, the Graduate School of Economic and Social Sciences, the Fontana Foundation, and the German Academic Exchange Service, things could have taken quite a different direction. I am very grateful to these four institutions. Similarly, I am thankful for the administrative support I was provided with. I think here of Sandro Holzheimer, CDSE center manager, Marion Lehnert, GESS team assistant, and Dagmar Rötsches, CDSE center manager. It was a real pleasure to interact with each of them and a source of comfort to know that procedural aspects would always fall into trustworthy hands.

I am deeply thankful to those who made it possible for me to turn to a happy private life the many times my work life took a difficult turn. I gladly conclude this section with a mention of them, starting with the truly fantastic people I was lucky to become acquainted with during the time I spent in Mannheim. Their strong and diverse personalities enriched mine in ways I would not have suspected. I am deeply grateful for every marks left upon me; they are a real source of joy. I think here of Alessandra Alloca, Albrecht Bohne, Francesco Paolo Conteduca, Robert Crowell, Majed Dodin, Andreas and Laura Dzemski, Thilo Eisele, Florian Exler, Verena Fetscher, Torben Fischer, Xin Gao, Timo Hoffman, Stefan Holand-Letz, Hala Jada, Ekaterina (Katia) Kazakova, Amélie Krauss, Vahe Krrikyan, Daria (Dasha) Khromenkova, Niccolò Lomys, Lonfei Lu, Maria (Masha) Marchenko, Kateryna (Kate) Miagka, Jan Sebastian Nimczik, Yasmin Özdemir, Marcel Preuß, Alexander Rholf, Christine Schnur, Barbara Singer, Chia-Yu (Joy) Tsai, Katarzyna (Kasia) Wróbel, Yihan Yan, Xue Zhang, and Maxim (Max) Zhilenkov.

As luck would have it, some among this lot have become close friends. With themselves, they brought a set of valuable and inspiring skills, which, to the extent I could learn them, proved really helpful to me. Oftentimes, they did so unintentionally, that is, just by being themselves. Kasia and Kate, for instance, instilled into me their resolutely positive outlook on life. In addition to making my own life more enjoyable, this gave me strength to leave worries outside the door whenever I heard them knocking. Dasha and Torben showed me, by being examples in this regard, that goal orientation is crucial to anyone who wishes to move forward and that, with reason and good will, one may reconcile pragmatic and humanistic values fairly well. Katia and Max (most friendly neighbors!) and Vahe marked me with their accute sense of solidarity (a form of life which, incidentally, fascinates my researcher-self). Finally, while Xue and Xin helped me to further develop my appreciation of simple pleasures, Marcel showed me that not so simple ones, too, have their importance.

Kasia, Kate, and Dasha went beyond impressing me with their personalities; They were the ones willing to become persons I could turn to at any time. Each managed this in spite

of my character. Kate and Dasha oftentimes made sure to cast a bright light on my person when it was difficult for me to do so. Kasia showed me her esteem, support, affection, and care in the most beautiful way: by embarking with me on a shared journey. Her daily company, she knows, makes me immune to adversity, thankful to life, and willing to move forward; I am happy she took this important place in my world.

My final thanks go to my family and longstanding friends. All granted me their unconditional support and affection. This, probably, hasn't been easy; because, in recent years, I made many questionable choices. I have only occasionally been attentive to them, to myself, and to our relationships; I have been insistent on following paths whose directions were unclear, to them and to me; I have been reluctant to communicate news; And I only seldomly took the time to head back home and spend quality time with them. For all that, my secondary school and later friends, Damien Arrestier, Antoine Béclin, Romain Blondet, Denis Debat, Audrey and Jean-Baptiste Gruson, Marjolaine Lecavelier, Christophe Lévêque, Céline Lighezzolo, Pauline Laugeais, Benoît and Clotilde Lopes, Stéphane and Émilie Ménoire, David Mirat, Jean-Marc Parveaux, Stéphanie Salidu, and Didace Sanou never failed to welcome me and share with me important moments and emotions occuring in their lives. For all that, my fantastic family, Yasmine, Bernard, Charlotte, Sophie, and Francois, never stopped expressing their belief in me, caring for me, and showing understanding and forgiveness for my absences or lack of involvement. The ability to unconditionally support is a wonderful, difficultly acquired skill. Certainly, I am very lucky that each of my siblings and parents masters it.

General Introduction

> "The fact that in common-sense thinking we take for granted our actual or potential knowledge of the meaning of human actions and their products, is, so I submit, precisely what social scientists want to express if they speak of understanding or *Verstehen* as a technique of dealing with human affairs. [...] It has nothing to do with introspection, it is a result of processes of learning or acculturation."
>
> **Alfred Schütz, Concept and Theory Formation in the Social Sciences, *The Journal of Philosophy*, 1954.**

A Linguistic Schism. At first sight, economists' scientific talk forms a single, homogeneous, form of discourse. All share a vocabulary: that of 'preferences,' 'beliefs,' 'choices,' 'decisions,' 'strategies,' 'uncertainty,' 'incentives,' and 'rational behavior.' And only a few raise concerns about the official selection of situations of greatest relevance for our field of study: 'equilibrium' situations and 'Pareto efficient' situations. It came as a surpise to me when, somewhat late into my studies, I realized that this homogeneity is largely a matter of appearances. Indeed, mainstream economists, despite using identical *words*, do not make use of identical *concepts*.[2] Rather, they divide themselves into two linguistic subcommunities (Sen, 1985): one which, one may say (Elster, 2009), is versed in *suspicion*, and another versed in *agnosticism*. When talking of 'preferences' or 'choice,' the former involve *volitional* concepts while the latter do not. When talking of 'beliefs' or 'uncertainty,' the former involve *perceptual* concepts while the latter do not. As a consequence, members of each respective community, when talking economics, do not wander in identical landscapes but in radically different ones. This is not without consequences, let us elaborate on this.

Economists versed in suspicion are more numerous than those versed in agnosticism. Accordingly, it is their phraseology that other social scientists identify with 'economic talk.' Possibly out of a concern for policy questions, though not necessarily, 'suspicious'

[2]The word 'concept,' itself, can be used to signify different things. To put it roughly, I identify "concepts" to *abilities*. The concept 'bird,' for instance, amounts to an ability to discriminate birds from other objects.

economists explicitly or implicitly take Hume's political maxim[3] to be applicable to economic matters. That is, they hold that, even though individuals eventually act on various motives, it is methodologically appropriate to assume that each is governed by her or his self-interest. A reason regularly appealed to is that preference maximization enables scientists to grasp an empirically relevant feature of human action; namely, that, often, its *efficient cause* is the fulfilment of desires (Elster, 2009). And three premisses recur in defenses of this line of thought. First, self-interested individuals may constitute, in an empirical sense, a majority among us. Second, there may be no reliable way to distinguish self-interested individuals from other kinds of individuals. Third, no technology may be available that would help us turn self-interested individuals into selfless ones. If each such premiss is true, then defenders of the suspicious approach dispose of a strong justification indeed.

Even though the first two premisses may not, formally, be contradicting each other, a tension does appear between them. If no trustworthy method exists that may help us distinguish self-interested individuals from those guided by other motives, what actually justifies our presumption that individuals are mostly self-interested?[4] Economists prone to raise this point form a minority; they are the *agnostics*. They take it that preference maximization has nothing to do with self-interest but only with rationality; that a rational actor must know what she wants and use the means conducive to her ends (Aumann (2000), Chapter 3). The most prominent agnostic was Samuelson; he held that, from the scientific point of view, the semantic content of a word like 'preference' ought to be confined to observable choice behavior (Samuelson, 1938, 1948). Now, clearly, this position runs against the intuitions we build up from everyday life. For, in everyday life, we *do* attribute to the word 'preference' a wider semantic content; possibly, one that involves psychological inclinations or even an intention. But a typical agnostic stands firm on her choice and give an argument for it. She says that it is not possible to have *objective knowledge* of the motives which direct human action. That the "knowledge" of someone's motives, one's own included, is based on introspection. And that even though, in everyday life, we take this "knowledge" for granted, scientists may not, because introspection is fallible and does not qualify as an objective source of knowledge.[5]

[3] "It is, therefore, a just *political* maxim, *that every man must be supposed a knave*; though, at the same time, it appears somewhat strange, that a maxim should be true in politics which is false *in fact*. But to satisfy us on this head, we may consider, that men are generally more honest in their private than in their public capacity." (Hume (1741), Part I, Section 6).

[4] Is it about generalizing from our introspective knowledge? Is introspection conducive to knowledge? Or, can one argue along Darwinian lines? Would such an argument be conclusive?

[5] Elster uses the conscious /subconscious distinction to build an interesting analysis of this fallibility. See, Elster (2009), Part I.

A Dispute of Methods. This lingusitic schism, I slowly got to understand, is not of little importance. It is a symptom of a long standing methodologial dispute between social scientists (see, e.g., Weber (1904)). On the one hand, proponents of an *interpretive* (*verstehende*) social science defend or take for granted scientists's ability to *understand* recurrent motives in social action and advocate, on this basis, a need for a methodological distinction between social and natural sciences. On the other hand, proponents of a *positivist* social science deny that social scientists have a possibility to do more than merely *explaining* (*erklären*) observable choice behavior in just the same way that natural scientists *explain* natural phenomena. The former (are committed to) agree with Schütz's contention that scientists can have insiders' knowledge about human behavior and that it makes sense for them to use this knowledge in their inquieries. They follow Weber (1904) in his conception of models as "analytical constructs (*Gedankenbild*)" within which the scientist gathers "certain traits, meaningful in their essential features, from the empirical reality of our culture." The latter, instead, (are committed to) refute this contention. They argue that, once wearing a scientist's glasses, the world you contemplate should be devoid of emotions, psychological drives, or motives; that it should only be populated with physical individuals, physical objects, and choice behavior; and that any other entity mentioned in our theories should be viewed as a mere heuristic device, a mean for us to summarize and convey information about the previously enumerated elements.

Each of the two camps, I believe, embodies a laudable set of concerns. To start with defenders of interpretive social sciences, they correctly see that it is hard, without some degree of *understanding*, to justify the use of established social scientific facts to *act upon the social world*. For instance, can we incentivize citizens to act responsibly? If so, should we and how? Is the issue at stake actually properly captured by 'incentive-talk'? Or should we rather use the classical frames of education and virtue? While understanding can provide guidance in these matters, a positivist claim to the effect that all we must do is change some parameters singled out as relevant by revealed preference analysis would be a qeer kind of claim. For one, it would run against important philosophical strands that lay foundations for Samuelson's view.[6] Second, it would entail a *conservative* stand that some among us will not feel obliged to accept unless they are given a solid reason for doing so. Now, conversely, it would be a clear breach to scientific ethic not to seriously consider the agnostic argument. Indeed, positivistically inclined scientists have long shown the value there is in epistemic conservatism.[7] Empirical evidence suggests that it plays an essential role when it comes to bringing about scientific progresses and that it sometimes prevents scientists from erecting epistemic obstacles (Daston and Galison,

[6] Wittgenstein (1921) argues against views that have facticity of description carry over to predictions.
[7] I.e., the tendency to raise questions related to the source of scientific knowledge.

2007).[8] So the question seems to come down to the following one: is the reason for dropping off understanding a '*solid*' one?

Chapter 1. The first essay I submit here is an attempt to answer this question. I observe that modern analytical philosophy gives us the means to (i) reject the conception of knowledge at the root of agnostic arguments, and (ii) defend Schütz's contention that *Verstehen* "has nothing to do with introspection, [that] it is a result of processes of learning or acculturation." My argument builds on essentially two pieces of philosophical work: Sellars (1956) and Anscombe (1957).[9] Sellars argues that knowledge is, before anything, a linguistic affair; that it is about developing a disposition to consciously conceptualize events in a way approved of by other members of one's epistemic community. Anscombe observed that (a) the concepts we mobilize when we identify events with instances of 'intentional actions' are not causal concepts but volitional and teleological ones; and that (b) when we ask an agent about her intention in doing something, what we are in fact asking for is not an avowal of her state of mind but information about *the* description under which she knows her own action. The agent's answer, as a consequence, isn't entierly private: it has to lie within a range of socially approved conceptualizations of the occuring events. Under this light, it appears that Schütz's processes of acculturation coincide with processes whereby individuals learn what appropriate descriptions exist for a class of specific events, viz. intentional actions. And knowledge of intentions is not merely "potential" but "actual:" it does not *essentially* differ from scientific knowledge. I conclude that, also in the event that introspection must be discarded from the scientific entreprise, it is not correct to conclude that we are left with choice behavior only.

Under the view developed in Chapter 1, what yields understanding is not introspection but the observation of specific forms of *linguistic behavior*. There, I do not identify linguistic behavior with individuals' description of their own actions but with the set of concepts at their disposal for describing events we refer to as intentional actions and, in particular, for regulating their or others' intentional actions. So conceptual analysis and folk psychology—branches of philosophy involved in the empirical analysis of action-, behavior-, and mind-related concepts—may be used, I claim, to enrich our analyses of economic phenomena. My two subsequent chapters are attempts to justify that claim.

[8] An epistemic obstacle is an unconscious structure present in scientific thinking which hinders advancements of science (Bachelard, 1938). Weber (1904) points, for instance, the common belief held by 19th-century scientists according to which a comprehensive rational system could be build that would encompass reality in it entirety.

[9] Both writers belong to a strand of philosophy quite foreign to Schütz's. I considered the possibility to delve into Schütz's own reasons for asserting this. I soon gave up. First, because I doubted I had time for the daunting investigations of continental philosophy it would have required. Second, because modern economics stands on building blocks laid down by analytical philosophers. A bridge with analytical philosophy would have been anyhow necessary.

Chapter 2. The second essay I present here is an illustration of the light that two regulative concepts can cast on phenomena involving a voluntarily provided public good. Namely, the concept of "joint commitment" (Gilbert, 1989) and that of "free riding." I draw a distinction between two kinds of public goods: ones I call *legitimate* and others I call non-legitimate. Legitimate public goods are particular in the sense that, in their presence, *individuals can use systems of demand-rights to check upon the behavior of their co-players*. For instance, abiding by a rule of etiquette is often tantamount to contributing to a public good. At any rate, it is so when the considered rule of etiquette promotes efficiency (e.g., observing silence in a classroom). And if you do not follow rules of etiquette while I do, then I have some authority to remind you of your engagements, I will exert that authority, and you will acknowledge it. But this does not happen in the presence of non-legitimate public goods, such as a crowdfunded Youtube broadcast or a crowdfunded charity. Indeed, being a contributor to either of these two goods does not endow me with an authority to demand from non-contributors that they make an effort. To take account of this discrepency, I suggest that, in the presence of a legitimate public good, individuals carry out a *joint intention to steer clear of free riding*. This is an assumption about equilibrium behavior, not about the structure of preferences. More specifically, in a standard Nash equilibrium, allocations ought to be stable with regard to potential individual deviations. But in the presence of regulative behavior, a deviation by one player may trigger the application of a system of demand-rights, whereby individuals who are now seen as making too low a contribution are authoritatively asked to increase their contributions. I design an equilibrium concept in which allocations are stable with respect individuals' deviations to the allocations they may reach by changing their contribution and exerting their demand-rights. I show that, within a specific class of demand-rights, the equilibrium that has all individuals contribute identical amounts is salient.

Chapter 3. The third essay I present here is meant to explore situations in which, in spite of strong reasons for making the public good a legitimate one, legitimacy is not yet established. For instance, when individuals are identically endowed, have identical preferences, and face identical production technologies, a commitment to steer clear of free riding does not merely help them deal with the free riding problem, it also helps them bring about a Pareto efficient situation. In such situations, it is fair to say of individuals that they have strong reasons for making the public good a legitimate one. Yet, they may not have had a possibility to do so. Experiments that give individuals an opportunity to volontarily and repeatedly contribute to a public good, I suggest, are instances of situations in which legitimacy could be of help but has not yet been established. Some well known stylized facts have been gathered about these experiments. For example, it was observed that contributions, although declining from period to period,

always remain significant. I investigate the possibility that this decreasing trend is the result of two phenomena: reputational concerns by strategic players and disagreement regarding the concept of "doing one's bit." Kreps et al. (1982) showed that, in finitely repeated public good games, reputational concerns may only come into play if some players behave irrationally or entertain a belief that some players behave irrationally. A first interesting finding I obtain in this chapter is this: their statement isn't robust to extentions of the concept of rationality. Indeed, I show that, when a subset of individuals is thought to abide by the principles of individual rationality and another by those of collective rationality, then there are instances in which it is individually rational to entertain a reputation of cooperator.

The thesis is structured as follows: the main findings of each essay and their analysis are presented in the respective Chapters 1 to 3. The first chapter's philosophical glossary as well as all the proofs for the two other chapters are respectively gathered in Appendices A to C. All due references are listed in the Bibliography, at the end of this volume.

Chapter 1

Mediated Preference Revelation

"On the most basic level, what we are trying to do in science is to understand the world. Predictions are an excellent means of testing our comprehension, and once we have the comprehension, applications are inevitable; but the basic aim of scientific activity remains comprehension itself."

Robert Aumann, *Collected Papers* (*Vol. 1*), 2000.

"I should explain that I am not arguing, in general, against internal correspondences that may be implied by the substantive nature of the specific exercise (e.g., by the maximization of an independently given utility function, if that is appropriate), but only against imposing such choice conditions in an *a priori* way as requirements of "internal consistency."

Amartya Sen, Internal Consistency of Choice, *Econometrica*, 1993.

1.1 Introduction

Many contemporary economists are positive scientists: they assert *facts*.*[1] To do so, they must learn to distinguish between those (sets of) statements which express facts and those which do not. It is conventional to teach them this skill not by the means of explicit discourses but by training them to *draw* the distinction (Kuhn, 1962). A neophyte economic modeler, for instance, will eventually have to meet the question "What does your model change?," usually meaning "What novel assertions are you making about choice behavior?" An explicit position regarding what facts are and how they have to be assessed

[1] Starred terms are terms I endow with a technical, sometimes idiosyncratic, meaning. Definitions are collected in Appendix A.

is called a *method*. A discussion advancing reasons for a method is called a *discourse on method*. The most notorious method in economics is Samuelson's theory of revealed preferences (Samuelson, 1938, 1948). The discourse on method that has had the greatest influence in its favor is Friedman's "methodology of positive economics" (Friedman, 1953). Revealed preference theorists assert (i) that ordinal utility theory and the principle of preference maximization—which, together, govern micro-founded economic models—should *not* be taken at face value, and (ii) that the preferences or beliefs they refer to are not real entities but *heuristic devices* whose unique purpose is to facilitate the transmission of information about economists' empirical data: individual *choice behavior*. They distinguish choice behavior from *intentional action*, for they claim never to assert any fact about intentions or mental states and they expect each of their theoretical statements to be translated into verifiable observation statements.[2]

In this paper, I am concerned primarily with the methodology of game theory. The first developments in game theory were not brought forth with positivist intentions. Rather, they embodied an attempt to delineate a *normative* account of rational action.[3] When game theory started to be used as an instrument for positive economics, disagreements appeared regarding its interpretation. For the sake of clarity, I shall classify these into two kinds. On the one hand, some asked after which kind of entities, present in alternative *forms of discourse*,* entities introduced in game theoretical discourse should be modeled. For instance, Rubinstein (1991) argues that strategies modeled after complete plans of action will trigger a different logic in scientific reasoning than strategies modeled after conjunctions of a plan of action and a set of justifications for that plan.[4] He also argued that thinking of a game form as a full description of physical events rather than as a comprehensive description of a rational perception of the situation will affect a scientist's modeling choices.[5] On the other hand, many took as given the set of entities present in game theoretical discourse (preferences, beliefs, actions, etc.) to instead raise questions about the desirability of postulating that such entitites, which are not *observables*,* actually exist. For instance, many practitioners implicitly commit, under the approving eye of philosophers of economics,[6] to the actual existence of preferences. But some prominent

[2] A short, immediate, illustration of the translation process is given by the meaning associated with the term "preferences". Let X be a set of available alternatives, B any subset of it, and $C^i(B)$ the choice set of an individual i presented with budget set B. When a follower of Samuelson asserts "agent i prefers x to y," he thereby means "Take $B \subseteq X$ such that $\{x, y\} \subseteq B$. If $y \in C^i(B)$, then $x \in C^i(B)$."

[3] On this point, see Aumann (2000), Chapter 3. A notable exception, it seems to me, is Luce and Raiffa (1957), which establishes a direct connection between ordinal utility theory and game theory.

[4] In the first case, it is natural to talk of a player's strategy 'choice.' But this is not true in the second case, where some elements of a player's strategy are thought of as being an opponent's belief.

[5] The first interpretation, for instance, entices us to take the notion of infinite repetition literally, but not the second one.

[6] See Hausman (1994) for a methodological argument, Okasha (2016) for a causal argument, and Dietrich and List (2016) for an ontological one.

theorists stand out in their defense of the view that, besides action profiles, game theoretical models merely consist of heuristic devices.[7]

The two sets of concerns differ in an important way. The first one relates to the delineation of a collection of entities, present in some other form of discourse, which it is best to formalize and integrate into game theoretical discourse. The second one takes as given the set of entities present in game theoretical discourse and asks which *epistemic attitude** is most appropriate for social scientists to adopt. On the second matter: one may be a (*scientific*) *realist*,* that is, commit to the actual existence of entities present in theoretical discourse, independently of their observability. Alternatively, one may be *anti-realist*, that is, commit to the actual existence of observable entities only. Each of the two attitudes suffers a downside. Anti-realism, on the one hand, must face the fact that policies based on mere metaphors can be harmful (see, e.g., Arendt (1972), Hayek (1975), or Ostrom (1990), Chapter 1). Successful alterations of an existing incentive structure, for instance, demands that agents *do* abide by preferences similar to those that economists take them to have. Realism, on the other hand, forces scientists to qualify as "knowledge" facts which are not outcomes of direct observation but outcomes of an inference to the best explanation.[8] There are known instances, in the history of science, when this attitude proved to be an obstacle to further advances in scientific knowledge (Daston and Galison, 2007). While it may well be that healthy dialectics between the two attitudes are more fruitful than the selection of either, this may only happen when each alternative is credible. In this chapter, I show that the case for realism has been, thus far, unduly weakened by two intricate misconception.

Call *behaviorist** the anti-realist attitudes akin to those of Samuelson and Friedman. There is a more popular,[9] realist, interpretation of economics, within which rationality is not identified with choice behavior but with the fulfillment of individual wants. More precisely, it isn't uncommon for economists to believe that preferences and beliefs are modeled after individual *mental states*, i.e., arrengements of internal physiological entities, and that, because mental states cause choice behavior, information about them may be inferred from actors' choices. Let me call *mentalist** any realist account of revealed preference theory in which it is contended that (i') individuals' mental life ought to explicitly enter economic modeling, (ii') whenever possible, observed choices ought to be *interpreted* as

[7] See Aumann (2000), Chapter 1, for a pragmatic argument, Gul and Pesendorfer (2008) for a methodological one, and Rubinstein (2012) for a rhetorical one.

[8] For a discussion of what "best" may mean in the context of economic explanations see, among others, Mäki (2006) and Sugden (2000).

[9] This point has been made by many philosophers of economics: among others, see McCloskey (1983), Hausman (2011), and Dietrich and List (2016). In fact, it is not at all clear that behaviorism, despite being the official standard, was ever held up by a majority of economists (Coase, 1982).

outcomes of an act of preference maximization, and (iii') preferences so revealed, call them *behavioral preferences*, actually coincide with individuals' mental preferences.[10] I claim (1) that the mentalist and behaviorist interpretations stand on two misconceptions and (2) that the mentalist view steepens rather than flatten, the hurdle faced by realism in economics. The latter blame is, in fact, quite transparent. Game theoretical models depend not only on preferences and beliefs, but also on their being commonly known. If mentalists have it right that preferences and beliefs are internal features of the actor's mind, then even the most standard common knowledge assumptions will prove very demanding. For they then raise questions regarding *how* individuals should ever reach such a level of shared knowledge about each others' mental states. I can see no reason to reject the behaviorist contention that no knowledge can be had of other minds; so long, that is to say, as one interprets it literally.

The actual focus of this essay is on the first claim, which is more subtle. Understanding and overcoming the aforementioned misconceptions, I argue, will allow us to reconcile our common knowledge assumptions with, on the one hand, the behaviorist contention that no knowledge can be had of other minds and, on the other hand, the mentalist contention that choice has to do with the fulfillment of wants. A first misconception, shared by behaviorists and mentalists alike, known as "belief in the Myth of a Given," is pointed out in Sellars (1956). To believe in the Myth of a Given consists in believing that observational knowledge is something that 'arises,' that is, that there exists a set of 'basic facts' whose truth or falsity may be merely observed by knowers. Instead, I will follow Sellars in his assertions that observational knowledge amounts to a conscious, rule governed, application of concepts to the witnessed events. From this perspective, mentalists appear to be asserting that, upon witnessing events we call 'intentional actions,' we conceptualize what occurs using physiological concepts. This, Anscombe (1957) pointed out, is a (second) misconception. Upon witnessing intentional actions we do *not* consider them *qua* outcomes of a causal, inner mechanism —even though they may be this, *too*. Rather, calling some event an 'intentional action' amounts to placing it in a teleological frame, to embedding it in a means-end relationship. Physiological and teleological talk, in turn, are radically different forms of discourse. The latter involves 'volitional' concepts and has us place events in a realm that goes *beyond* that of physics. For, unlike the former, it aims not at *explanation* of human behavior, but at its *regulation*.[11]

[10] Not all writers insisting on the need to import vocabulary about mental states into game theoretical discourse take a realist stand. Rubinstein and Salant (2008), for instance, assert that "there is no escape from including mental entities, such as the way in which an individual perceives the objects and his mental preferences, in economic models." Yet, in other writings, Rubinstein is quite clear about his reluctance to accept a realist point of view (Rubinstein, 2012). Views akin to Rubinstein's are, as far as I can see it, instances of behaviorism.

[11] A simple thought experiment illustrates our linguistic awareness of the cause vs. reason distinction.

A consequence is this: upon acting rationally, i.e., upon trying to bring about events which observers could describe as rational actions, an agent *must* embed her action in a series of means-end relationships that observers would recognize as having application in the situation she finds herself in. Now, reasons for acting, the elements of discourse that enter those means-end relationships, *are* observables. We observe them and get to know about them by engaging or witnessing *regulative* talk about human behavior. I draw two conclusions from this. First, Sen's contention that preference maximization need not be seen as an *a priori* methodological principle *does* have foundations. Indeed, we can conceive of it as an *empirically relevant rationale.* Second, to reach an *understanding* of a situation, economists should widen their evidence base to include, in addition to choice behavior, pieces of *linguistic behavior* relevant to that situation. Different situations are observed to kindle approval for the use of different regulative concepts and rational agents, upon framing a situation, distinguish between concepts that have application and those that do not. For instance, it was long recognized by Smith (1776) that the concept of *benevolence* has no application in situations of trade. Or, as I will argue in the next chapter, the concepts of *obligation* and *free-riding* have application in the presence of *some* public goods but not others. Scientists willing to understand behavior in the presence of the first kind of public goods must recognize the fact that, in their presence, preference revelation will be *mediated* by these two regulative concepts.

The remainder of this chapter is structured as follows. I start with a brief overview of two mainstream views of the logic of scientific discourse: the mentalist and behaviorist one. The behaviorist philosophy of science, it is well known, strongly influenced mid-century economic methodology. As such, it constitutes a potential reason why many economists show some reluctance to accept the identification of preferences with mental states. In section 3, I discuss some limitations of the behaviorist philosophy of knowledge and present Sellars' alternative. Sellars' view brings no support to the identification of preferences to mental states. But it sets the ground for Anscombe's account of intentions. I detail the latter, and emphasize that Anscombian intentions, as opposed to mentalist ones, are not entirely private. Indeed, they neither are properties of the actor's mind nor properties of the observed events, but consitutents of the from of discourse that we use to conceptualize events we call 'intentional actions.' In section 4, I detail the implications of Sellars and Anscombe views for scientists' conception of rationality. In particular, I argue that it brings us closer to solving the puzzle of common knowledge faced by the mentalist approach. I briefly conclude in section 5.

Saying of a person that she is a *cause* of your happiness is not the same as saying that she is a *reason* for your happiness. The easiest way to check on this is to substitute happiness with despair. A person may be a cause of your despair, but she may not be a reason for you to despair. Despair is difficulty conceived of as (an intermediate step for) a good. More on the distinction in sections 1.3.2 and 1.4.1.

1.2 Economics and The Logic of Scientific Discourse

Discourse—communication in speech or writing—is not a single, unified *practice*,* but a collection of practices (Wittgenstein, 1953; Austin, 1979). I will call discursive practices *forms of discourse*. Each form of discourse is characterized by a unique set of *rules* that legitimates moves from some *inputs*—features of reality, elements of perception, or previously uttered sentences—to *outputs*: a sentence or a set of sentences. For instance, *logical discourse* is a practice which consists in applying rules of inference—e.g., the modus ponens—to sets of propositions in order to yield still further propositions. Upon engaging in a logical argument, a speaker's perfomance will be correct or mistaken according to whether or not, each times she makes a move, she does so in conformity with one of the pre-specified rules of inference. I call the set of rules that characterizes a form of discourse its *logic*.* Consider the two following questions, which have relevance for our topic. First, does going beyond *as if* preferences *necessarily* thwart the objectivity of the social sciences? Second, in the event that it need not, does mentalism constitute an adequate alternative to behaviorism or is another interpretation needed? I want to argue that our disagreements with regard to these questions find their roots in disagreements about the rules that govern scientific discourse.

Rational individuals willing to make factual assertions cannot do so arbitrarily. They ought to operate in ways which respect the standards of science, i.e., they must engage in scientific discourse. Behaviorists identify scientific discourse with two practices: theoretical and observational discourse. Theoretical discourse, they say, is the practice in which individuals engage when they *assert true analytic statements** (or deny false ones). Observational discourse is the practice in which individuals engage when they *assert observation statements* (or deny false ones), i.e., statements such as "this object is blue," or "individual i chooses product A." The mentalist view can be seen as equally separating scientific discourse into theoretical and observational practices (Hume (1739) and Hayek (1937) do so), but the rules they associate with each activity are radically different. Theoretical discourse, they say, is the activity in which individuals engage when they *report relations of ideas*, and observational discourse the one in which individuals *report matters of facts*. An important difference is that mentalist reports, which may be truthful or untruthful, are governed by principles of ethics, while behaviorist assertions, which may be adequate or mistaken, are governed by skill. A further difference concerns the domain of objects over which a knower's moves are defined. Mentalists assert that the inputs of observation and theoretical statements alike pertain to the private realm of individual perception. Behaviorists hold that each kind of statement takes as inputs elements in the realm of *public** physical entities. I must expand a little on this.

1.2.1 Two Logics for Scientific Discourse

The mentalist syllogism[12] starts with an account of what is *given* to us through *perception*. Call *sense-impressions* the inner experiences to which an individual is subject when confronted with her environment. Mentalists draw a distinction between sense-impressions and *mental images*, which they define as faint but accurate copies of the sense-impressions. When an individual experiences a sense-impression, her mind systematically forms a mental image of that impression. Given these, mentalists reason as follows. First, they identify *reasoning* with a series of movements whereby an individual brings different mental images in the presence of one another and judges whether these are in some way *related*. Second, they classify relations between images into two kinds. *Relations of ideas* are relations between images which we may know to be *necessarily* true or false because their assessability follows from the involved images *only* (e.g., 'two is larger than one'). *Matters of fact*, in contrast, are relations between images whose truth-value has a *contingent* character. Beyond reasoning, their assessment demands observations, i.e., supposed (but unjustifiable) correspondences between simple impressions and the outer world. An individual may report about the relations of ideas she has disconfirmed or confirmed; such reports constitute pieces of theoretical discourse. A report about matters of facts, in contrast, constitutes a piece of observational discourse.

Hume's writings, in which the distinction between relations of ideas and matters of fact was first drawn (see Hume (1739), Book I, Part III, Sections I and II), constitute a representative instance of the mentalist view. Hume counts as relations of ideas all logical relations, such as those of ordering or negation, as well as all conceptual relations, such as that between a triangle and the value of the sum of its angles. He includes in potential matters of facts relations of resemblance, of contiguity in space and time, and of causality. Finally, he takes the presence of two types of relations to yield two possible forms of knowledge. When an individual assents to a mental image she experiences because, using relations of ideas only, she is able to relate it to self-evident relations between simpler images, then her knowledge is *a priori* and it inherits the *necessary* character of relations of ideas. If, instead, the assent to an experienced image requires the use of a matter of fact, then her knowledge is *a posteriori* and it inherits the *contingent*, less secure, character of observations.[13] The representativity of Hume's account holds in

[12] My reconstructions of the mentalist and behaviorist syllogisms, naturally, are very rough. The very suggestion that there should be 'a' mentalist and 'a' behaviorist way to look at things makes it clear. I do not and cannot seek accuracy here. My less ambitious intention is to briefly outline two logics which I shall later criticise. Hopefully, the critique is robust to fine grained variations in the various positions that mentalist or behaviorist authors actually held.

[13] Indeed, *a posteriori* truth involves a *given*, an unjustifiable claim of correspondence between reality and sense-impressions.

the following sense: to mentalists, knowing amounts to experiencing a specific *mental state*. Discourse may be had about facts, but it is logically preceded by the experience of knowledge and it consists in the utterance of intentional reports about the contents of one's mind.

Behaviorists rely on a very different syllogism: one associated with a philosophical movement known under the name of *logical positivism*.[14] At the core of their argument stand three ideas. First, that observational discourse applies to physical reality directly and not, as mentalists would have it, indirectly through mental perceptions. Second, that theoretical discourse does not consist of reports about cognitive self-evidence, but in the unfolding of *analytic* truths, i.e., truths which we uncover by following conventional rules of language. And third, that scientific discourse logically precedes, as opposed to being logically preceded by, knowledge. To set things straight, let us call a *conceivable state of affairs* any singular arrangement of physical objects. A *realized state of affairs*, then, is an arrangement of physical objects that actually obtains. Logical positivists define reality as a partition of conceivable states of affairs into realized and unrealized ones. They further distinguish *meanigful* sentences, i.e., sentences that express a statement to which the concept of truth may be applied, into two kinds. There are, on the one hand, synthetic sentences, which express a *structured proposition*, i.e., a representation, in symbols, of a conceivable state of affairs. Their truth or falsity is *a posteriori*. And there are, on the other hand, analytic sentences, which express statements whose truth or falsity obtains by virtue of the (definitional) meaning of the symbols they contain and, hence, is *a priori*.

Their line of reasoning runs as follows. Call *observation reports* those sentences which express a *basic proposition*, i.e., a symbolic depiction of a state of affairs so simple that its truth can be assessed ostensively without being doubted ("Here, now, light."). The utterance of an observation report, behaviorists contend, is the mere expression of a disposition humans have to ostensively learn language. As a consequence, observational discourse is the outcome of a series of trained dispositions whereby individuals become able to utter, in the presence of a simple state of affairs, observation reports. It does not involve a knower's ability to produce mental images in any *essential* way.[15] Coming

[14] The account is akin to Ayer's *verificationism* (Ayer, 1946), Hempel's *physicalism* (Hempel, 1935), and what Samuelson (1948) refers to as *operationalism*.

[15] A good illustration occurs in Wittgenstein (1953), §6. "[The] ostensive teaching of words can be said to establish an association between the word and the thing. But what does this mean? Well, it may mean various things; but one very likely thinks first of all that a picture of the object comes before the child's mind when it hears the word. [...] But if the ostensive teaching has this effect,—am I to say that it effects an understanding of the word? Don't you understand the call "Slab!" if you act upon it in such-and-such a way?" What Wittgenstein points at here is this: think of a child deprived of mental images and nontheless able to pick the slab anytime her teacher calls "Slab!" Would we not be willing to say that the child understands the word? If so, then the mental image cannot be what we refer to when we mention the 'meaning' of the word "Slab." And it is in no way *essential* to the ostensive learning process.

now to theoretical discourse, let *truth functions* be functions whose inputs and outputs are propositions and whose output has a truth value determined by that of its inputs.[16] Logical positivists follow Wittgenstein (1921) in his assertions to the effect that all rules of meaningful language are truth functions (see, esp. §6). This position, eventually, allow them to equally reduce theoretical discourse to a series of trained dispositions; dispositions whereby individuals become able to apply *truth functions* to propositions. On this view, theoretical discourse, just like observational discourse, is devoid of substantive mental contents, and the truth of all scientific sentences may be reduced to a series of behavioral episodes. Knowledge of a structured proposition is achieved by identifying basic propositions of which it is a truth function and ostensively verifying these basic propositions. Knowledge of an analytic statement is achieved by identifying simple tautologies (e.g., the laws of excluded middle and of noncontradiction) of which it is a truth function.

To sum up, mentalists and behaviorists offer two different accounts of knowledge. These have in common the recognition of two types of knowledge: *a posteriori* knowledge, whose justification has to do with facts, and *a priori* knowledge, whose justification is independent of facts. But they differ in that the former identify knowledge with a specific type of mental state while the latter identify it with behavioral episodes. Note that *a posteriori* knowledge is the most relevant type of knowledge for the topic of this essay. For, if knowledge of other minds is possible, as mentalists contend, then it is a type of knowledge that depends on actually realized states of affairs.

	Mentalism	**Behaviorism**
Discursive domain	Mental images	Physical Reality
Knowledge		
Truth	*Correspondence* (between sense-impressions and events)	*Verification* (ostensive assessment of basic propositions)
Justification	*Foundationalism* (Direct or indirect relation to true sense impressions)	*Foundationalism* (Truth function of true basic proposition)

Table 1.1. Table 1. Mentalist vs. Behaviorist *a posteriori* Justified True Belief

In the two subsequent subsections, I detail further the logical positivist account and show its connection with the foundations of neo-classical economics.

[16] An example is logical conjunction, the truth value of proposition $p \wedge q$ is fully determined by that of p and q, as the associated truth table shows. Truth tables are credited to Wittgenstein and Post.

1.2.2 Structured Propositions

The contention that *a posteriori* knowledge may be characterized without making a reference to the knower's mental states is not an intuitive one. Ordinary ideas about knowledge are closer to the mentalist syllogism: they involve mental states, oftentimes a form of enlightenment.[17] Nonetheless, logical positivists did manage to give convincing arguments against this claim. It will take this subsection and the begining of the next one to see how. Let us start with the distinction they draw between a '*sentence*' and the '*statement*' it expresses. A sentence's statement is its semantic content, i.e., the information that is being conveyed from one individual to another when both understand the sequence of symbols that make up a sentence. One can easily distinguish it from the sentence by using synonymous sentences in different languages: "Socrates is a man," "Sokrates ist ein Mann," and "Socrate est un homme," are three sentences which express the same statement. Logical positivists folow Frege (1892), Russell (1905, 1919) and Wittgenstein (1921) in their identification of empirical statements with *structured propositions*, i.e., symbolic representations of reality. Wittgenstein puts it this way: "in order to understand the essential nature of a proposition, we should consider hieroglyphic script, which *depicts* the facts that it describes" (§4.016, emphasis added). The central, unintuitive, contention they add to this view is that, to explain how we come to associate states of affairs with their respective structured propositions, there is no need to make any reference to mental states.

I already mentionned that logical positivists identify *reality* with the existence and non-existence of *states of affairs* (Wittgenstein (1921), §2.06). A structured proposition, then, is a representation in language of a state of affairs. At first sight, it seems that, even for a sentence as simple as "Socrates is a man," our grasping of the expressed state of affairs *does* involve mental states in a rather essential way. To see why it *need not* be so and, therefore, why it may be possible to build a scientific language that does not depend on them, let us first ask ourselves: what state of affairs is being represented by the proposition expressed in "Socrates is a man"? Intuitively, one may want to say that the proper name refers to an existent entity—*Socrates*, that 'a man' equally does so (just as it seems to do in the sentence "a man crosses the street") and that 'is' refers to an identity relation that holds between the entity represented by 'Socrates' and that represented by 'a man.' But this view cannot be correct. For, if it were, 'a man' and 'Socrates' would refer to the same individual, and asserting "Socrates is a man," would merely amount

[17] We easily picture someone uttering "Eureka!" or "I know!" as being in one way or another enlightened. Following Descartes, we take it that "God as given us an inner light to distinguish the true from the false"(Descartes (1637), p. 24) and attach these mental phenomena to the essence of what it is for someone to know something.

to asserting a tautology: the identity of an individual with himself. Clearly, that is *not* the meaning we associate to that sentence. This suggests that the relationship between sentences and the structured propositions they express is not a straightforward one. And this implies that, to be in a position to emit a judgment regarding the necessity, or its absence, of a reference to mental states of the discursive knower, we must be better informed about the rules of meaningful language, i.e., the general features of structured propositions.

The clarification of the relationship between sentences and their expressed propositions is to be credited, to a large extent, to Frege, Russell, and Wittgenstein. Frege (1892) identified the kind of issues I presented in the previous paragraph. Russell (1905, 1919) came up with the suggestion that one may deal with sentences akin to "Socrates is a man" by identifying indefinite descriptions (here: 'a man') to instances of unitary relations. In particular, we ought to think of 'a man' as a relation of inclusion in the set of all men, whatever this, in turn, should mean.[18] But many more questions remains. For instance, "Socrates is a man," once analysed along Russelian lines, still involves entities we call 'relations.' And since a proposition, on Wittgenstein's own words, is like a "hieroglyphic script" in that it "depicts the facts that it describes," it now seems that Russell's view commits us to the assertion that 'relations' we depict are objects as real as, say, *this* apple. This was, in fact, Russell's stand (see, e.g., Russell (1912), chapter IX). But not all philosophers readily accept such commitments, so Russells view could not have imposed itself without significant modifications. It is Wittgenstein (1921) who, upon trying to generalize Russell's theory of descriptions, suggested a more acceptable variant of it. He argues against Russell's contention that relations, like names, are depicting terms.[19] For if it were true, an infinite regress would arise in that one could still ask: how does the depicted relation, *qua* existing object, relate to the other objects? Rather than being depictable real entities, relations are elements of punctuation in the proposition. They endow propositions with a *structure*, which, in turn, enable us to give them their sense, i.e., to associate them with a corresponding states of affairs.

Wittgenstein's contentions about real entities did not stop there. Eventually, he argues, *all real entities must be of a single kind*. He calls them "objects" or "simples" and conceptualizes them as black shapes on a white surface. From the logical standpoint, he observes, an object must be characterized by its *form*, the collection of the possible ways in which it

[18]The cautious reader will have justified worries about the status of 'a man' in the sentence "a man crosses the street." Russell's suggestion is to reformulate such sentences in the following way: "There exists x, x is a man, and x crosses the street."

[19]"Instead of "The complex sign 'aRb,' says that a stands to be b in the relation R,"' we ought to put "That 'a' stands to 'b' in a certain relation says that 'aRb."' (§3.1432), "Situations can be described, but not given names. (Names are like points; propositions like arrows—they have sense.)" (§3.144)

may combine with other objects (§2.011-2.01231), rather than by any kind of *content*. As a consequence, objects are not further analyzable (§2.02) and 'names' are the symbols we use to depict them, the symbols that "*stand for*" them (§3.203, §3.22-3.221). States of affairs, in turn, are equally identifiied by their *form* (or *structure*): a unique way in which the objects it involves are combined. If we consider the *possibility* that such or such structure may exist, we express what he calls a *logical picture* of that state of affairs, also known as, a (structured) *proposition* (§3).[20] Accordingly, language, the realm of structured propositions, is the collection of pictures that truth functions entitle us to form about reality (§3.03). It is an *a priori* realm (§2.222-2.225, §4.51, §5.61) in the sense that it's elements are neither true nor false but only contain the possibility of truth or falsity (§3.13). To declare an empirical fact true or false, in turn, is to hold up its associated logical picture, correlate it with reality, and assert "This obtains" or "This does not obtain" (§2.201).

Propositions, so described, relate to reality in a specific way: using language is projecting reality onto a subset of it, the set of symbols (§3.1-3.141). Meaning is preserved if the projection follows the rules of meaningful language, that is, if the proposition is a *truth function* of elementary propositions, where elementary propositions are depictions of reality so simple that we may straightforwardly correlate them with reality. Every proposition, then contains up to two kinds of constituents: *logical constants*, punctuation-marks that encode the way in which reality is projected onto the space of 'hieroglyphs' (§5.4611 and §6.124), and *names*, the 'hieroglyphs' proper. Elementary propositions correlate straightforwardly because they are mere concatenations of names, i.e., they do not involve any punctuation sign (§4.22, see also Anscombe (1971)). Much of Wittgenstein's work consists in giving credit to the view that, eventually, all genuine synthetic statements may be reduced to a set of elementary proposition to which a series of truth functions have been applied (§6). This doctrine, despite its abstractness, should ring a bell to economists familiar with Samuelson's claim that statements involving theoretical terms like 'preference' ought to be reduced to specified observable phenomena: episodes of choice behavior. Eventually, a correlation between propositions and reality consists in the identification of each name contained in the proposition with an object of reality, and in this identification only (§5.4733). Questions regarding the nature of this identification gave rise to logical positivism, to which I now turn.

[20] I here conflate Wittgenstein's propositions and proto-propositions. Given the aim of the paper, I take it that it is better so.

1.2.3 Logical Positivism and Behaviorist Economics

The logical positivist movement was initated by a group of Viennese philosophers and scientists who, in the 20's and 30's, held regular meetings during which topics in philosophy of science were discussed. The group, which became known under the name 'Vienna Circle,' developped ways of thinking about knowledge that greatly influenced mid-century science, behaviorist economics included. Logical positivists took over much of Wittgenstein's account of language (see, e.g., Cavaillès (1935)) and adjoined to it a theory of knowledge, i.e., an account of the conditions under which the truth value of structured propositions comes to be ascertained. The logical positivist theory of knowledge is embodied in three main variants: *verificationism*, *physicalism*, and *operationalism*. Proponents of either variants accept Wittgenstein's identification of reality to a partition of conceivable states of affairs into existent and non-existent ones. They also accept his contention that states of affairs eventually amount to singular combinations of a unique type of entity, the 'simples.' And, importantly, they turn the claim that all structured propositions are truth functions of elementary propositions into a *foundationalist* thesis. That is, they claim that the edifice of knowledge stands, as an upside down pyramid would, on a set of propositions we reach certainty about: the elementary propositions. Knowledge of elementary propositions, in turn, they take to be non-inferential, i.e., to stand on its own feet. Schlick (1934), for instance, has it that statements, *a priori* and *a posteriori* alike, which may be non-inferentially known all share a common property. Namely, that one may not *understand their meaning* without simultaneously *assessing their truth-value*.

Consider some analytical sentence, say, one that expresses a logical truth. It's incontrovertibility arises from the fact that *understanding its meaning*, the expressed proposition, and *assessing it's truth-value* are one and the same (noncognitive) event.[21] If we look at synthetic statements, matters appear, at first, to be radically different. Understanding the proposition expressed by a synthetic statement amounts to grasping which observable events would put an observer in a position to assess the truth of that statement. Put differently, it amounts to identifying the entities that one has to correlate with each name involved in the proposition in order to be able to say "This obtains," or 'This does not obtain." And certainly, understanding so conceived need *not* amount to an *actual* verification of the involved facts. Nonetheless, consider a situation in which an individual develops a disposition of the following kind: she can utter statements of the

[21] Tautologies and contradictions provide good illustrations of this assertion. Understanding the proposition codified by "$p \vee \neg p$" and assessing its validity are one and the same event. Furthermore, it is difficult to characterize this event as a "cognitive" one, since it just amounts to a correct use of the disjunction.

form "Here, now, so-and-so" when, and only when, specific states of affairs arise. For instance, she may have an ability to utter "Here, now, apple, falls," or "Here, now, x, y, choice of x" when and only when the specified events occur. Such linguistic abilities—call them utterances of *observation statements*—are like a pointing at observed states of affairs.[22] Upon executing them, we "carry out the process which is necessary for the verification of all [empirical] statements" (Schlick (1934), p. 225), i.e., we necessarily verify their truth. Observation statements, in other words, constitute a class of synthetic statements for which understanding and truth-value assessment are coincident and non-cognitive.

Going back to Wittgenstein's account of language and substituting Schlick's non-inferential knowings for his elementary propositions, we obtain an account of scientific knowledge as exposed in Ayer (1946).[23] On that account, a known statement is a statement that expresses a proposition which is a logical construct of non-inferential knowings: tautologies or observation statements. Scientists remain free to introduce theoretical terms into their discourses, but only in the sense that one remains free to name p the logical construct $a \vee b \wedge c$, where a, b, and c are basic propositions. Theoretical terms facilitate the scientific enterprise by simplifying the encoding of empirical observations. But it is a mistake to give them a meaning other than that of being logical constructs of elementary propositions, for this is the only objective meaning they can have. This view, when connected with economic analysis, yields interesting implications. For instance, since neither preferences nor beliefs may refer to observable, physical entities, their use in economic models must be that of theoretical terms. And their meaning must be arising from a possibility to translate them (without remainder) into observation statements. This is precisely the "as if" doctrine. I conclude this section with pieces of evidence against the coincidental nature of this connection.

During the 30's most members of the Vienna Circle emigrated to the U.S., and ther is evidence that their ideas raised the interest of intellectual leaders in some major universities. The Boston area, where Samuelson obtained his degrees and started his career, is known for having welcomed much of the diaspora.[24] So Samuelson's account of revealed preferences, which advocates the substitution of a 'technical' concept of preference for the "discredited" psychological one (Samuelson, 1938), need not come as a surprise.

[22] A child who has just learned the word "car" and repeats it everytime s/he sees one could be seen as doing just that kind of gesture. See also footnote 15.

[23] In the first edition of his book, Ayer rejects Schlick's account of observation statements (see, esp. pp. 90-91). But he corrects himself on that point in the introduction to the second edition (see pp. 10-11).

[24] An account of the "Vienna Circle in Exile" that settled in Boston is given in Holton (1995). A welcome can signal intellectual affinities. Indeed, Bridgman, who's ideas on science are explicitly referred to by Samuelson, is a co-founder of the Unity of Science Institute, along with Quine, Carnap, and Frank.

Further evidence can be found in two forms. First, direct connections between individuals can be established. Ramsey, who suggested to substitute a 'technical' concept of belief for the psychological one (Ramsey, 1931), was well acquainted with Wittgenstein. In fact, the former is known to have spent much time discussing with the latter the accuracy and consequences of his *Tractatus*.[25] Similarly, Morgenstern, who moved from Vienna to the U.S. in the late 30's, acknowledged having "struggled hard with Ludwig Wittgenstein's *Tractatus*" and "frequently" attended meetings of the Vienna Circle (Morgenstern, 1976). Second, the terminology used in mid-century economic writings is revealing. Savage and Arrow, for instance, respectively assert that 'neo-Bernoullians' "improve on Bernoulli in that [they] define utility *operationally* in terms of the behavior of a person constrained by certain postulates" (see Savage (1972), section 5.6, emphasis added) and that "the *only* meaning the concept of utility can be said to have is their indications of actual behavior" (Arrow (1963), p. 9, emphasis added).

Beyond Samuelson's explicit attempts to "[drop] off the last vestiges of the utility analysis" (see Samuelson (1938), p. 62) and to construct "*operationally* meaningful" propositions (expression repeatedly used in Samuelson (1948), emphasis added), defenders of a more pragmatic line of reasoning could be met too. Such thinkers found sufficient to lay the emphasis on the *possibility* to proceed without taking any stand regarding the actual existence of mental states. Friedman, for instance, draws a semantic distinction between statements about unobservables entities, referred to as *hypotheses*, and statements about observables, referred to as *consequences*, and argues that *there is no need* to read hypotheses literally, that their only purpose is to help us to organise assertions about consequences. He asserts that "viewed as a language, theory has no substantive content;" that "it is a set of tautologies;" and that "its function is to serve as a filing system for organizing empirical material and facilitating our understanding of it." (Friedman (1953), p. 148.) Luce and Raiffa upon asserting that "there is no need to assume, or to philosophize about, the existence of an underlying subjective utility function" (see Luce and Raiffa (1957), section 2.6), seem to take a similar stand. This view did not satisfy Samuelson, for whom Friedman's argument induces the reader to think about unobservable entities as really existent but merely dismissed from scientific discourse.[26] It nontheless remain influential, as the following statement found in Aumann (1998) shows: "To avoid mis-

[25] "Since I began to occupy myself with philosophy again, sixteen years ago, I could not but recognize grave mistakes in what I set out in that first book. I was helped to realize these mistakes—to a degree which I myself am hardly able to estimate—by the criticism which my ideas encountered from Frank Ramsey, with whom I discussed them in innumerable conversations during the last two years of his life" (see Wittgenstein (1953), Preface).

[26] See Samuelson (1963), where the author complains about the logical inconsistency of the 'F-Twist' and its propensity to induce the reader to believe that abstract theoretical terms could be taken to refer to real, unobservable, entities.

understanding, we stress that we do not consider the CPA [Common Prior Assumption] 'true;' the concept of truth does not apply here. We do think that it embodies a reasonable and useful approach to interactive decision problems" (p. 929).

1.3 The Undermining of Realism in Economics

"Scientific realism is a positive epistemic attitude toward the content of our best theories and models, recommending belief in both observable and unobservable aspects of the world described by the sciences" (Chakravartty, 2017). The credibility of realism in modern economics, possibly, is undermined by significant adherence to two erroneous beliefs. First, many influential economic theorists[27] use a rethoric reminiscent of the logical positivist and instrumentalists conceptions of theoretical discourse. They take it that, in economic models, preferences and beliefs are theoretical terms, that theoretical terms do not name any causal explanans, and that therefore one would be mistaken when attempting to correlate them to whatever common-sense preferences and beliefs are names for. Albeit right on some aspects, this conception misses out the explanatory role of theoretical discourse and rests on a radical and problematic distinction between scientific and common-sense knowledge. Second, many who do not adhere to the anti-realist view just mentioned hold that econmists' basic unobservable entities (preferences, beliefs, etc.) are refinments of homonymous common-sense concepts: concepts from folk-psychology. Again, there is something to it, but the folk psychology that we refer to must be rightly construed. In this section, I take up each issue in turn.

1.3.1 The Normativity of Observational Knowledge

A first hurdle for defenders of logical positivism comes from casual empirical evidence: their account of knowledge seems to be used and endorsed by only a minority among practicioners of science.[28] As far as economists are concerned, it has been argued that their rhetoric is based on wider grounds than those pushed for by logical positivist, and rightly so (see McCloskey (1983)). Or, that many among them do not look for mere *explanation* (*Erklärung*), but for *understanding* (*Verstehen*) (Coase, 1982; Sugden, 2000). In a nutshell, this says that, when it comes to selecting among competing theories, economists do not exclusively value empirical adequacy but operate a trade off between

[27] I already mentionned Aumann (1998), see also Gul and Pesendorfer (2008) and Rubinstein (2012).

[28] Empirical evidence for the physical sciences is provided in Bachelard (1938), Kuhn (1962), and Polanyi (1958). Empirical evidence for economics is discussed in Coase (1982) and McCloskey (1983). Finally, Goffman (1959) discusses the case of everyday life.

empirical adequacy and alternative criteria exploiting their common-sense knowledge of social life. This casual empirical hurdle, however, is not the only one. Indeed, from a purely fromal perspective too, logical positivism, in the form in which it was presented, faces difficultes. For instance, Anscombe (1971) (see esp. Chapter 1) shows that the identification of Wittgenstein's elementary propositions with observation statements is unfaithful to Wittgenstein's own account of them. I now present in greater details two formal issues that "any historian of [the shift from logical positivism to later analytical philosophy] would do well to focus on" (Rorty, introduction to Sellars (1956)).

For one, Quine (1951) argues, it is not logically possible to combine, as logical positivists do, the claim that the meaningful entities of a language are its sentences and the claim that a language derives its meaning from a potential correspondence with states of affairs. The reason has two sides. On the one hand, the claim that individual sentences are the meaningful elements of a language goes hand in hand with an important distinction: that drawn between analytical and synthetic statements. As it happens, the only unequivocal way to maintain a sensible version of this distinction, i.e., one that would include some extra-logical synonym pairs, is to use the verificationist account of synonymy. That is, to say of two statements that they are synonymous "if and only if they are alike in point of method of empirical confirmation or infirmation" (p. 35). On the other hand, it is quite clear that, independently of how minute one is, what is being verified is never a single proposition expressed by one sentence but—and inescapably so—a collection of propositions expressed by a collection of sentences, a *theory*. The logical positivist identification of the foundations of knowledge with analytical and observation statements—for which it is true to say that understanding and verification amount to the same thing— cannot reconcile these two sides. Although verificationist-synonymy may happen between theories, it cannot happen between individual sentences.

Sellars (1956) brings another, decisive, argument into the debate. The logical positivist account of knowledge rests, we have seen, on a reduction of meaning to the depiction of physical states of affairs and, eventually, on a possibility for individuals to develop a disposition to 'correctly' point at simple, realized states of affairs. The idea of 'correctness' involved here, he underlines, is understood by logical positivists as the correctness of a body movement. But that, clearly, cannot do. For, if the relationship of observation statements to the occurring events really is of this kind, then observation statements are a mere continuation of particular sensations. *As such*, they *cannot* logically entail knowledge. Knowledge is knowledge of *facts*, not of particulars (§3). A fact, when properly asserted, is endowed with *normative authority*, that is, it is expected from the knower that she can give a appropriate justification for her assertion. The ability to

justify, in turn, signals that the knower is following rules of language *consciously*, not merely 'correctly' (§34, §35). Sellars, who sides with logical positivists in their reluctance to refer to mental states, points out that reference to inner perceptions will not do as a justification (§38). His original move is to put to work a distinction duly noted by Rawls (1955) around the same time: that between justifying a practice and justifying a particular action that fall under this practice. The correctness of fact stating must be construed, he argues, "as being an instance of a general mode of behavior which, in a given linguistic community, it is reasonable to sanction and support" (§35). In other words, the correctness of observation statements is justified by reference to an existing practice.[29]

The point found in Sellars (1956) bares resemblence with the one made by the later Wittgenstein (see, Wittgenstein (1953), esp. §1 - 15). Namely, that facts do not bear out the long held belief that language learning reduces to ostensive learning, and, in particular, that it does not demand any sort of reference to a context. Meaning arises from and is learned from the use we make of language *in specific contexts*. The forms of discourse have their elements—words, sentence tokens, or collections of sentences—used in accordance with rules of language whose shape originates in the end that gave the practice its use (see Wittgenstein (1953), §11, for an imaged expression of this point). Sellars further explores Wittgenstein's idea (see, esp. §29-30) and argues that the authority of observations statements originates not in a *given*, an unjustifiable correspondence between word and world, but in the recognition that, within the context of "looks talk," it is appropriate to single out this observation report as a "reliable symptom" of some realized fact. Without such a recognition by the knower, no actual correlation of words with world can be said to have taken place (Sellars (1956), §33 - §38). An economics professor discloses factual knowledge when, upon observing an individual's choices, she asserts that the individual prefers apples to pomegranates. A fresh student who repeats the professor's assertion while thinking about the pleasure of bitting apples and the burden of reaching the seeds of a pomegranate, does not.

The kind of recognition we just mentioned, clearly, presupposes knowledge of a general fact of the kind "*X is a reliable symptom of Y*." This is a severe issue for logical positivists, for it implies that amending their account in a way that accomodates the concept of a 'correct' observation report will not do. Their foundationalist doctrine then leads them into an infinite regress: observation of particular facts requires the use of general facts,

[29]The reference need not be explicit. For instance, Sellars entices us to think of justifying observational knowledge by naming further observable objects or properties. In this way one implicitly shows that the circumstances are "normal" and that there are no reasons for doubting "looks talk." Theoretical assertions, differently, may be justified by reference to the facts that they are usually taken to be causally explaning.

which, in turn should be motivated by particular facts, and so on. Sellars, who gives up on foundationalism, is not concerned with this issue. Knowledge, on his account, amount not to tracing back one's statement to a class of foundational, verifiable, non-inferential knowings. Rather, knowing a fact amounts to "being able to justify what one says," that is, to being able to refer to a discursive practice in which the concerned statement is a factual statement. We may come back to our fresh economics student who, after a combination of mistaken assertions and reminders about Samuelson's concept of revealed preferences, reaches a stage where, after having observed an individual's behavior, she can assert of him that he prefers a to b and justify her assertion.

It is important not to leap into the belief that, by rejecting foundationalism, Sellars lets anything go. X counts as "a reliable symptom of Y" *in the context of a practice*. That is, the general fact tho which one may refer in order to justify one's knowledge of particular facts must have been singled out by some practice. And the authority with which this general fact will be endowed will have to do with the empirical relevance and rational appeal of the said practice. Thus, in line with Quine's defense of holistic empiricism, Sellars simply appears to shift the burden of truth from isolated statements to a set of scientific discursive practices. And even though it is well known that the priviledged relationship between truth and scientific discursive practices still is an ongoing debate among philosophers of science,[30] acknowledging this fact does not amount to a claim against science. "Empirical knowledge, like its sophisticated extension, science, Sellars writes, is rational, not because it has a foundation but because it is a self-correcting enterprise which can put any claim in jeopardy, though not all at once" (Sellars (1956), §38). This view of science, importantly, opens up a road for Sellars' further assertions about folk psychology. If scientific discourse combines, to various extents, different forms of discourse, then why should we not think that this is equally true of common-sense discourse? Towards the end of his essay, Sellars suggests that common-sense discourse does entail some amount of theoretical talk. He adds that a neat way to understand mental states may be to construe them as theoretical entities which individuals would have taken a habit to refer to when looking for causes of behavior. Here are the beginings of folk psychology as a theory (Ravenscroft, 2019).

1.3.2 Intentions as Forward-Looking, Volitional Explanantia

It is beyond doubt that our mental experiences may affect our acceptance of beliefs. It is the case, for instance, when we are subject to wishful thinking. But wishful thinking,

[30] Even when it comes to the most established scientific discursive practices, such as logic and mathematics. See Carroll (1895), Quine (1948), and Maddy (2012).

precisely, is defined *in opposition to* authoritative acceptation of beliefs; it may be cited as a *cause* of our belief but it may not amount to having a *justification* for accepting this belief. This observation can be cast into an interesting light: upon asking a person 'why?' an individual does this or that (e.g., why she accepts a belief?), we may not be asking for a *causal explanans*, an entirely private kind of explanans, but for some other, not entirely private, kind of explanans. For, we have just seen that, on Sellars' view, rational acceptation of belief gets its justification from a *normative explanans*, a reference to the rules of a practice. What about instances of the question 'why?' that aim at eliciting an actor's *intention*? Intentions, which are central to any attempt to *understand* social life, were banned from the scientific realm by mid-twenteeth century methodological changes in economics. The stipulated reason was, on Samuelson (1938)'s words, to "drop-off the last vestiges" of utility as a "discredited psychological concept." In other words, Samuelson asserted that, since economists may never get clear about the causal explanans of individuals' actions, they should restrict their observation statements to statements about choice behavior. That actions may be linked to normative explanantia already casts doubt on the the validity of Samuelson's syllogism. In this section, I follow Anscombe (1957)'s argument to the effect that a consistent account of intention can neither take them to be causal nor normative explantia, but may take them to be forward-looking, volitional explanentia.

Today, many a scholar still has it that folk talk about intentions really is causal-talk and that it refers to mental states (Scheer, 2004). But this view, Anscombe (1957) argued, is mistaken and owes its prominance to essentially two facts, which any account of intention ought to accomodate. For one, it is indeed the case that, whenever one inquires about an individual's intention in doing this or that, one is always bound to leave the last word to the actor herself (§4). For instance, if I ask you about your intentions in writing this letter, and if your answer has the appropriate form, I have no choice but to accept your authority on that matter. Call this fact the existence of a *first-person authority*. Second, it is also the case that, upon acting intentionally, an actor seems to *groundlessly* know the intention with which she is acting. For instance, consider a situation in which you want to open your office window. Upon standing up from your chair, no observable event indicates whether you are about to pay a visit to the neighbor office or to open the window. Yet, *you* know you are opening the window. Call such cases of knowing cases of *groundless knowledge*. The presence of this specific form of knowledge gives comfort to the idea that, somehow, you must be 'directly observing' mental states of yours, i.e., observing them in a way different from the way we observe outer events (§29 and §32). Goundless knowledge and first person authority, Anscombe argues, "conspire to make us think that if we want to know a man's intentions it is into the content of his mind, and

only into these, that we must enquire" (§4).

But it is not the case that all evidence is in favor of the mentalist account of intentions either. First, expressions of intention, in most cases, describe a state of affairs that is yet to come. If, as mentalists view it, they were avowals of a current state of mind, then an explantion should be given for the existence of a causal connection between such states of the actor's mind and the futur state of affairs that is being mentioned (§2). It is difficult to believe that a state of mind can intelligibly cause or bring about the occurence of a determinate state of affairs. Presuming that this state of mind is called to mind by the actor will not help. For then the calling to mind would itself be an intentional action; an infinite regress would be entered (§19). Second, we can make an additional observation about expressions of intention, Anscombe puts it this way: "It is not the case that a description of *any* future state of affairs can be an answer to [a question to a man about his intention]. A man's intention in acting is not so private and interior a thing that he has *absolute* authority in saying what it is—as he has absolute authority in saying what he dreamt" (§22, second emphasis added). Indeed, if, upon being asked about her intentions in doing this or that, an agent gives an arbitrary answer, then she runs a risk of not making sense. She may not for instance, give a mere physical description of what she is currently doing—the questioner, who *sees* that, does not ask for it. Nor can she claim to be bringing about a state of affairs which by no means will be brought about by the kind of movements she is currently performing.

A single intellectual move, Anscombe suggests, enables us to get in line with all of the four empirical observations we just mentionned. She invites us (§2 and §32) to pay greater attention to differences in the way an individual *knows* what *causes* her action and the way she *knows* the *intention* with which she acted. In particular, we should look at the way error about causes differ from errors about own intentions. If a desire to eat chocolate is what causes me to hold a belief that having 80g of chocolate a day is a healthy habit, I still may believe that I hold this belief because I feel healthier since I took this habit. In this event the mistake I make is a *mistake of judgment*: my belief about what causes me to believe in the healthy features of chocolate is wrong because *it does not match the events*. Consider now the case in which I accepted that belief with an intention to bring my eating habits in line with ministerial recommendations. It may be that I am mistaken, i.e., that this belief is not in line with ministerial recommendations. But in that case, the mistake I make is not one of judgment. Rather, I wrongly infered, maybe, from the recommendation that a daily intake of fruits is healthy, and the fact that cacao is a fruit that I should hold the belief that a daily intake of chocolate, whose main ingredient is cacao, is healthy. This is a *mistake in performance*: the events are not matching

the description under which I know my belief acceptance; not the other way around.

This observation is key in the following sense: it is only because we have a tendency to identify all forms of knowledge with *contemptlative* (*a.k.a. theoretical*) *knowledge* that we feel concerned with the groundlessness of our knowledge of the reasons with which we act. "If there are two knowledges—one by observation, the other in intention—then it looks as if there must be two objects of knowledge; but if one says the objects are the same, one looks hopelessly for the different *mode of contemplative knowledge* in acting, as if there were a very queer and special sort of seeing eye in the middle of the acting" (§32). But nothing prevents us from asserting that knowledge of ones own intentions is a *practical* kind of knowledge. That is, we may assume that agents, upon acting intentionally, use conventional forms of descriptions of events *as blueprints* for their own actions and try to bring about events in conformity with this form of description. This would explain groundlessness, partial first-person authority, and avoid us the trouble of making unjustifiable causal claims. For, within a linguistic community, rational individuals may not describe events *arbitrarily* but must abide by *justifiyable forms of descriptions*. And although the relation between events and forms of description need not be bijective, the realm of possible intentions in doing this or that remains finite and, importantly, a possible object of contemplative knowldge.[31] When an agent acts *rationally*, the first-person authority comes only to settle possibly arising issues about (i) intentionality in acting, and (ii) identification of *the* description under which her actions are intentional. A rational individual, certainly, is not granted the right to pick any description she pleases.

As for what intentionality in acting is, Anscombe suggests that "we do not add anything to the action at the time it is done by describing it as intentional" (§19). For, to call an action intentional simply is to place it in a conceptual frame that goes "beyond physics," a conceptual frame centered on the concepts of life and animality (§47). This, I believe, can be understood along Sellarian lines.[32] Knowing is about placing events in a conventional conceptual framework and, upon being asked, being able to justify one's placing by a reference to that convention. Anscombe is arguing that intentional actions simply are a subset of events which we know not under the concept of physical causality but under that of volition.[33] Volitional explanantia are a forward-looking kind of explanantia, they

[31] Being disruptive in the way one looks at things is a sort of event that equally possesses its form of description, i.e., conceptualization. Namely, being disruptive.

[32] And maybe also by more continental approaches to philosophy. For, on this point, Anscombe and Sellars seems to me to line up with Sartre's contention that "the world of explanations and reasons is *not* that of existence" (Sartre (1938), p. 148), translation and emphasis are mine.

[33] "Consider a question 'What is the stove doing?', with the answer 'Burning well' and a question 'What is Smith doing?' with the answer 'Resting'. Would not a parallel answer about Smith really be 'breathing steadily' or perhaps 'lying extended on a bed'? Someone who was struck by this might think it remarkable that the same expression 'What is—doing?' should be understood in such different ways: here is the case of

relate to the state of affairs that the individual wants to bring about. As such, they neither coincide with the normative kind of explanans that Sellars mentions—these are looking back at a practice—but constitute a third kind of action-explanans, the second that isn't entirely private. Now, either of (i) a brief recollection of the Ancient Greeks' convention to place movements of birds into the sphere of divine intentionality or (ii) a thorough historical study of the formation of the concept of a 'reflex' (Canguilhem, 1955) shows that the set of events which we place into one or the other of the three categories changes as time elapses. But the time of universal determinism, certainly, has not yet come.

1.4 Mediated Preference Revelation

It is now time to take stock and draw some of the conclusions that follow from our previous arguments. There are essentially two. First, contrary to Samuelson (1938, 1948)'s assetions, it is not the case that dropping the psychological concept of utility leaves economists with choice behavior *only*. Folk psychology hasn't only developped along a *causal*, mentalist dimension, but along a *regulative*, mind-independent dimension too McGeer (2007). Individual action, with which social scientists are concerned, to the extent that it is rational, *must* meet the standards of folk psychology as a regulative practice. That is, rational action *must* be action performed in view of making a description of the state of affairs come true, and it *must* be based on a justifiable conceptualization of the occuring events. Therefore, independently of our ability, qua scientists, to know the *causal*-explanentia of observed behavior, we still have a possibility to know their *volitional*- and *normative*-explanantia, because these, on top of being recorded in the regulative dimension of folk-psychology, stand on their own feet. Second, the conceptual framework in which we place events we call "intentional actions" has an *observable* structure. One which, actually, is carefully described in Anscombe (1957). She showed that descriptions of intentional actions relates in an essential way to practical reasonings. And this suggests that, between internal consistency of choice and desire fulfillment, an alternative view of rationality can be drawn. One in which a rational action is a form of behavior practically known to the agent under a description that (i) meets the linguistic standards of the agent's linguistic community, and (ii) may be viewed as the starting point of a *valid* practical reasoning.

---

the 'enormously complicated tacit conventions' that accompany our understanding of ordinary language, as Wittgenstein said in the *Tractatus*," (Anscombe (1957), §43).

1.4.1 Rational Behavior and Regulative Forms of Discourse

When it comes to suggesting an alternative to behaviorism, the flag of mentalism is traditionally held up in the economic literature. I identify mentalism with three theses: (i') conformity (in some form) of an individual's rational behavior with his experienced mental states, (ii') interpretation, whenever possible, of observed choices as outcomes of an *actual* act of preference maximization, and (iii') agreement of behavioral (revealed) preferences with the individual's actual (mental) preferences.[34] Mentalism, so characterized, is a realist position which, I have already argued, raises an epistemic problem that is quite independent of whether or not one adheres to Sellars' view. Game theorists model interactive situations in which individual beliefs or preferences are, completely or incompletely, commonly known. It is unclear how mentalists can justify that agents dispose of such an extensive knowledge of other minds, or even that scientists get to know facts about these. In this section, I argue that, because rational action *is* behavior in conformity with regulative talk about behavior, and because the empirically relevant concepts of regulative talk are observables, an alternative to mentalism exists which does not face such a severe epistemic problem.

We have seen that, additionally to forcing social scientists to base their policy recommendations on mere metaphors, the behaviorist approach rests on a questionable philosophy of knowledge. These arn't the only issues it faces. Sen vehemently criticized it for (i) its failure to take into account committed behavior (Sen, 1977), (ii) the unreasonable demands which its main assumption, the revelation assumption, imposes on individual choices, be it under risk (Sen, 1985) or in a certain environment (Sen, 1973), and (iii) its lack of logical consistency (Sen, 1993). There is a difficulty, however, with some of the arguments typically brought against behaviorism. For instance, consider the following thought experiment (from Sen (1985), p. 110):

> *"Take a choice function $C(\cdot)$, assumed to be 'rationalizable' (i.e., 'binary') and let R be the binary relation representing it. Construct the binary relation R^* from R by 'reversing' every strict preference, and let $C^*(\cdot)$ be the choice function generated by (and 'rationalizable' with respect to) R^*. If a person with unchanged non-choice characteristics (i.e., the same feelings, values, tastes, etc.) were to end up choosing*

[34] One could be tempted to identify mentalism with (i') and (ii') only, and thus to reject (iii'). In fact, this seems to be the view taken in some strands of the litterature, for instance in research on learning in games (see, e.g., Fudenberg and Levine (1998) or Hart and Mas-Colell (2013), in particular the explicit discussion of 'uncoupledness' in latter). Under this view, while mental life matters, knowledge of it remains private and the hypothesis of preference maximization merely registers the fact that individuals tend to do what they want, without ever asserting anything about the objects of their wants. But it leads to a dead-end: if knowledge of others' wants is impossible, then Nash equilibrium is not within reach (Hart and Mansour, 2010; Hart, 2011).

> *in exactly the 'opposite' way in each case, i.e., according to $C^*(\cdot)$ rather than $C(\cdot)$, it would be hard to claim that his or her choices have remained just as 'rational'. But the 'opposite' choices are exactly as consistent!"*

There is a sense in which this argument will appear convincing to a mentalist but *cannot* convince true behaviorists. For, we have seen it, behaviorists dismiss the very possibility to make factual assertions about objects or qualities mentalits call "values," "feelings," or "tastes." A consequence is that, without an account of *observable non-choice characteristics*, Sen's argument is bound to prompt the "no-nonsense operationalist[35]" to (mischievously) ask: 'What do you *mean* by values and feelings? How, in point of fact, does the one differ from the other?[36]"

Combining Sellars and Anscombe's views solves precisely *this* problem. Sellars tells us that *knowing* amounts to framing witnessed events in a way that is approved of in the context of some practice. Anscombe (1957) tells us that "intentional actions" should not be seen as actions fulfilling some specific properties (say, consistency), nor as actions performed by an individual whose mind is in a specific state, but as events in individual's life that *we* frame using specific concepts. Namely, events for which our framing process makes use of 'animist' or 'vitalist' concepts. Furthermore, she gives a detailed analysis of the specific form that description of intentional actions may take. "Aristotle's 'practical reasoning,' she says, [..] can be looked at as a device which reveals the order that there is in [the ordinary language conventions that regulate intentional concepts]" (§43). There is but one way to understand her analysis: intentional actions are events which can legitimately be identified with a conclusion in a piece of practical reasoning. And while this would be of no avail if Hume had it right that when being practically rational we always aim to fulfill our desires (see Hume (1739), Book II, Section III, Part III), Anscombe dismantles this view too (§30). Practical reasons have to do with states of affairs *becoming* true, as opposed to states of affairs *being true*. They involve *knowledge* of the eventuality as well as an *attempt to bring it about*.[37]

It is worth insisting on that last point. The belief that desires and interest lay down the law about a man's wants is a widely accepted one, and its effect is to push social scientists in either of the mentalist or behaviorist camp. For, it *is* the identification of practical rationality with the fulfillment of desires that drives policy concerned scientists

[35] Sen's nickname for behaviorists, see Sen (1985), p. 112.

[36] Ayer (1946), an important proponent of logical positivism indeed developed an own account of a view called *emotivism*. Emotivists identify ethical and value judgments to expressions of feelings.

[37] Anscombe's reasoning on that point is fairly abstract. It is not useful to replicate it here. Consider instead the following intuition, suggested in Foot (2001): "Many of us are willing to reject a 'present desire' theory of reasons for action because we think that someone who knowingly puts his future health at risk for a trivial pleasure is behaving foolishly" (p. 63).

into an endless search for universal laws of self-interest; and it *is* this same identification that give a reason to epistemically conservative scientists to conlcude that a technical concept of rationality, based on consistency of choice, is necessary to conduct objective scientific analysis. By pursuing a line of reasoning that undermines this belief, Anscombe brings grist to an interpretation of game theory that differs from the two mainstream ones: Bacharach's rational approach (Bacharach, 2006). Bacharach insists that the frames within which individuals reason, too, are objects to be studied by scientists. As he puts it (p. 7),

> *"people evidently do reason, more or less well... Moreover, some of the reasons that plausibly guide people's behaviour are very general and so have great explanatory power—for example, the reason for choosing an alternative that it maximises expected utility."*

His second point is of great importance: validity of the reasoning is one thing, but empirical relevance matters more. Clearly so for social scientists, who are concerned with that very matter. But *also* for actors of everyday life, who, in interactive situations, must pay a cost for abiding by valid but empirically marginal practical reasonings. This may be seen from a life-experience narrated by Rousseau (1782) in his sixth walk. There, Rousseau is concerned with the possibility for him to keep his 'freedom' and perform 'good deeds.' Roughly put, his reasoning is as follows:

1. A good action is an action carried out with an intention to do bring pleasure to another person;
2. A man's freedom consist in having a choice not to to act against his desires; [38]
3. Whenever I (Rousseau) perform a good action, my intentions are "misjudged" by my beneficiaries; they take me to be acting virtuously;
4. Acting virtuously and performing a good action are two different things. A virtuous action is an action carried out in order to play one's part in a 'society' that benefactor and beneficiary form together;
5. Virtuous actions entail duties which, once I recognize them, anihilate my other-regarding feelings and cause me displeasure;

⇒ "The only good in my power from now on is to refrain from doing anything for fear of unintentionally and unwittingly doing ill."

[38] I take it to be the case that Rousseau had a desire-fulfillment theory of individual welfare. There is also room for a Kantian reading of his conception of liberty (see, e.g., Rousseau (1762), Book I, Chapter VIII). The matter is not essential to the general argument.

I do not think that Rousseau's example, in which the idiosyncratic character of his framing of events makes it impossible for him to bring about the said events, presents us with an isolated case. Quite the contrary, I believe that this squares well with insights we may gain from a great variety of strands in the literature.[39] It is in line with the intuition, captured in Plato's allegory of the cave (Plato (380), VII 514 a, 2 to 517 a, 7), that disruptive modes of reasoning tend to create barriers in an individual's social life. At any rate, little evidence suffices to suggest an interesting possibility. Namely, that the chances for a mode of reasoning to become empirically relevant in a given interactive situation depends on the success it brings to the individual (or groups of individuals) who adopt it. *Individual* rationality, on this view, is neither an *a priori* methodological principle, nor an *a posteriori* and sovereign principle of human nature, but a mode of reasoning that yields strong reasons for acting in many economically relevant situations. Of course, we also know of some situations in which individual rationality yields only very weak reasons for acting. It is the case when each among the many can choose to pay a price and cast a vote, or when it comes to providing a public good. In such situations, other modes of reasoning, maybe a rule-utilitarian one (Harsanyi, 1977b), a reciprocal one (Sugden, 1984), or a Kantian one (Roemer, 2010, 2015), yield stronger reasons for acting.

1.4.2 Solving Common Knowledge Issues: the Case of Bilateral Trade

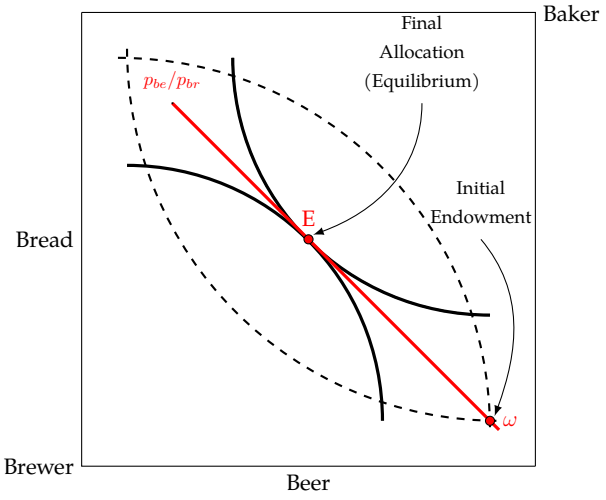
In the first chapter of the *Wealth of Nations*, Smith (1776) suggests that the division of labor is the most important factor of economic development. In the second chapter, he seeks to find a "principle which gives occasion to the division of labor." His suggestion is that "the propensity to truck, barter, and exchange one thing for another," which "is common to all men," constitutes such a principle. It is clear that he took this statement quite literally.[40] Similarly, many modern economists have no troubles accepting Smith's talk of propensities. But not all. Those with behaviorist inclinations, for the reason we have just seen, take it that any such principle should be translated, without remainder, into observable choice behavior. This divergence in 'opinions,' eventually, has consequences for scientific practice. While mentalists can freely talk of incentive based policy, behaviorist can only difficultly do so. In this section, I detail what facts are being reported,

[39] In the realm of the social sciences, I can think of Goffman's assertions to the effect that individuals "project" a "definition" of the situation and that these projections "limit what it is the individual *can* be," (Goffman, 1959) and of Bourdieu's concept of a 'field.' The 20th century literature, too, is rich in characters with idiosyncratic, inflexible modes of reasoning who, much like Rousseau, end up socially isolated or ill appreciated. Meursault (Camus, 1942), Clappique (Malraux, 1933), Morel (Gary, 1956), and Hans Schnier (Böll, 1963) all are instances.

[40] On this matter, the analytical introduction to Smith's Wealth of Nation by Andrew Skinner is clear. Skinner identifies Smith as a member of the Scottish Enlightenment, a school of thought which asserted the existence of propensities in human nature "independently of our knowledge of them." (see, pp. 12,13)

upon analyzing bilateral trade, by each of the two schools. I also illustrate how, even in the event that mentalist assertions fail to convince, there *is* an alternative to behaviorism that provides grounds for incentive based policies.

On behaviorists' view, talks of "propensities to trade" are metaphorical talks, that is, mere linguistic devices helpful to keep track of, or convey information about observable physical regularities. The expression can be, and scientists ought to *reduce* it to an expression involving observable choice behavior only. In the present case, the observation that individuals often operate simultaneous bilateral choices, whereby goods of one type are exchanged for goods of another type at a certain rate. In this way, assuming a "propensity to truck, barter, and exchange one thing for another" does not commit the scientist to any factual claim about the actual mental states of the trading partners or about their intentions. It merely presents us with a concise way to summarize assertions about the allocation of goods among individuals and individuals' propensity to bring about changes in that realm.



- Facts can be stated about:

  Initial endowments,
  Final allocations,
  The exchange rate.

- Heuristic device:

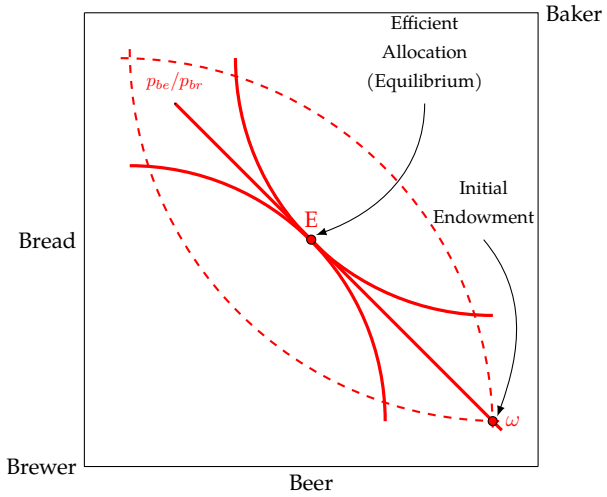
  Preferences.

Statements about preferences *summarize* gathered information about initial endowments, final allocations, and exchange rates.

On the behaviorist account, equilibria do not record patterns of individual mental states or perceptions of the situation but patterns of individual behavior. There can be no scientific ground for restrictions on the shape of indifference curves other than those entailed by choice observations. For all we know, it could be that the indifference curves pointing upwards are those of the Baker—who wants, say, to have the brewer fed and talkative. Similarly, the ones pointing downwards could be those of the Brewer.

It is clear that the behaviorist interpretation fails to give a ground for implementing incentive based policies. Without knowing what the baker and brewer *want* we cannot give them incentives to alter their behavior. This yields a pragmatic reason for economists to commit to the existence of some mental states, among which knowledge and belief, as

well as to the existence of mental entities, such as preferences. On this view, preferences are *causal* determinants of an individual's actions. Observation of behavior enables us to draw inferences about them, but introspection constitutes an alternative source of information; it gives us an opportunity to state facts about indifference curves *also* in the absence of behavioral observations. Smith himself must be viewed as taking such a position. For instance, he takes it for granted that each trader seeks to maximize the amount of beer and bread in his hands, and that, for this reason, the indifference curves pointing upwards must be those of the Brewer and those pointing downwards must be those of the Baker.



- Facts can be stated about:
  Initial endowments,
  Final allocations,
  The exchange rate,
  Mental states.

Statements about preferences *record* information about an (ideal) individual's mental states.

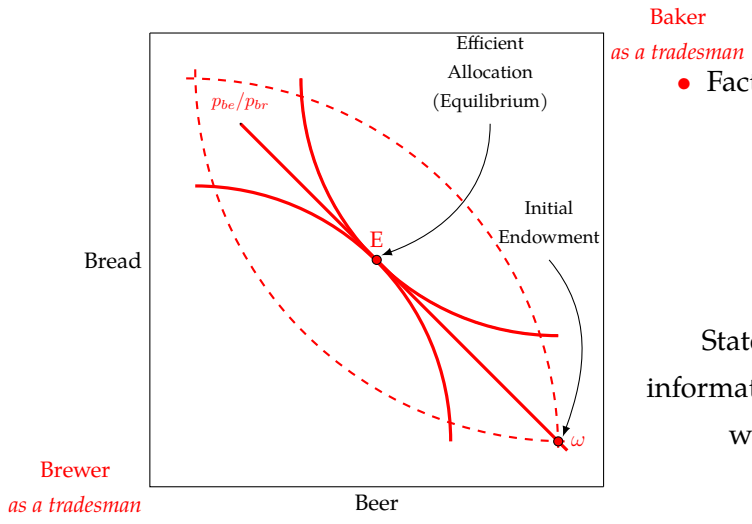
Mentalists work with ideal-types, that is, theoretical constructs summarizing factual information about properties which are thought to be widely shared by real individuals.[41] Thus, they interpret statements about preferences as statements which they expect to be approximately true—maybe in a statistical sense—or credible about individuals' real preferences. In the present case, a desire to possess larger stocks of goods, possibly formalized with an assumption of monotonous preferences. Such statements could, in principle, be justified by the means of behavioral observations only, but *de facto* they arn't. Mentalists rhetoric, instead, relies on alternative principles such as, for instance, similarity principles regarding *human nature* (Harsanyi, 1977a).

Neither of the two mainstream views, I argued in this essay, is convincing. Behaviorism, with its world of allocations and exchange rates, has an "air of paradox" (see Ayer (1946), p. 20). We do seem to know more about each other's actions than the way in which they affect the allocation of goods in society. At the same time, if it is agreed that introspection

[41] The theoretical nature of ideal-types does not go against realism. Realism does not require exact correspondence between words and world. Example of ideal-types of preferences in economics are selfish preferences, risk averse preferences, other-regarding preferences, or distributional preference.

can difficultly be counted as a legitimate source of scientific knowledge, then mentalism fails to tell us what *additional information* we have about individuals' behavior and *where from* we have it.[42] The alternative I advocate is in fact suggested by Smith's own choice of words. He holds that our disposition for barter is more likely to be a consequence "of the faculties of reason and speech" and, on several occasions, identifies trade with a form of "contract." "Two greyhounds, he writes, in running down the same hare, have sometimes the *appearance* of acting in some sort of concert. Each turns her towards his companion, or endeavors to intercept her when his companion turns her towards himself. This, however, is not the effect of any *contract*, but of the accidental concurrence of their passions in the same object at that particular time. Nobody ever saw a dog make a fair and deliberate exchange of one bone for another with another dog." (Emphasis added, p. 118).

I want to suggest that Smiths' observations need not so much be about human nature as about *the nature of trade*, about trade as an established *practice* with an (implicit) contract. What we call trading isn't a mere exchange of good, but the act of exchanging goods *with an intention to get the best material bargain for oneself.* Individuals who master the *concept* of trade *know* this. The baker need not *explicitly* tell himself that, upon engaging in a trade with the brewer, he will seek a good bargain; yet, because he *practically knows* his intention to "trade," i.e., because he tries to bring about a state of affairs that an observer who masters the concept of trade would find fit to describe as an token of trade, he also *knows* the situation is also about getting a good bargain, he knows that the brewer, because he, *too*, masters the concept of trade, knows it is about getting a good bargain, and he knows the brewer knows that he knows, *ad infinitum*.



- Facts may be stated about:
  Initial endowments,
  Final allocations,
  The exchange rate ,
  The practice of trade.

Statements about preferences *record* information about a *practice*, that of trade, in which individuals may engage.

[42] One possible road might be the theory-theory of folk psychology. But it is necessary to note that, along this line, a significant difference exists between 'anonymous' interactive situations—i.e., interactive situations involving many individuals and in which actual physical interaction are reduced to a minimum—and simpler situations in which a few individuals are in the presence of one another. Much of our 'explicative' folk-psychological knowledge relies on direct observation of body movements.

Each society, eventually, is characterized by a set of social practices in which its individuals regularly engage. Information about these practices is recorded in the concepts under which they are subsumed and conveyed by the means of *regulative* talk. Individuals who master these concepts can rationally use them as blueprints for their own actions. Events we describe as tokens of "trade" *cannot* be described as tokens of "charitable giving", tokens of "coerced exchange," or tokens of "symbolic exchange." Each of these concepts cover a different ground and individuals mastering them know what each entails. The activity of trade, for instance, involves a "contract" whose terms are recorded in the concept of trade.[43] This contract, Anscombe (1958) shows, involves a concept of *obligation*—an 'owe;' but, as correctly pointed out by Smith (1776), it does not involve the concept of *benevolence*. Each society, eventually, bears the responsibility of the practices and concepts it puts forward. Recent developments have shown that the concept of trade can be amended and that it may, for instance, come to involve ideas of distributive justice. Individuals now have a possibility to choose between engaging in a classical trade or in a 'fair trade.' Upon choosing the one over the other, they know which attitude to expect from their trade partners, which attitude their trade partners expect from them, they know that their partners know the same, and that they know that they know, *ad infinitum*.

1.5 Conclusion

One of Samuelson's reasons for holding up the flag of revealed preference was this: it is not possible to acquire knowledge about another individual's mental states, therefore, social scientists are bound to use choice behavior as the *only* building block of their models (Samuelson, 1938). In this chapter, I did not contest the premiss in Samuelson's argument. Rather, I argued against its conclusion. Even in the event that no knowledge of mental states can be gained, we are not justified in saying that, to keep on with the epistemic demands of science, social scientists ought to reduce rational behavior to consistent choice behavior. The reason is twofold. First, the behaviorist contention that knowledge may be reduced to a series of behavioral episodes has long lost its appeal among philosophers of knowledge. Sellars (1956) showed that this account rests on a myth, the 'Myth of the Given,' and that it fails to explain the *normative* nature of knowledge. The normative nature of knowledge, he argues, stems from the ability to relate

[43] With regard to the parallel between trade and contracts, the following statement from the *Theory of Moral Sentiments* is illustrative: "[The individual]...in the race for wealth and honors and preferments... may run as hard as he can, and strain every nerve and every muscle, in order to outstrip all his competitors. But if he should hustle or throw down any of them, the indulgence of the spectators is entirely at an end. It is a violation of fair play, which they cannot admit of." (II.ii.2.1), quoted in Skinner's analytical introduction.

one's factual statements to an established discursive practice. Knowing consists in placing events in an inferential argument, and because this 'placement' is subject to approval or disapproval by our peers, knowledge has a normative status. Second, and contrary to a belief shared—for different reasons—by mentalists and behaviorists, 'intentions' and 'mental states' relate to 'intentional actions' in radically different ways. While 'mental states,' independently of how we conceive of them, take the role of causal explanans for our actions, Anscombe (1957) showed that 'intentions' are nothing but redescriptions of our actions. Redescriptions in regulative terms.

There are, I believe, three important consequences that follow from the argument I just developped. First, in order to give a realistic turn to our assumptions about rational behavior, economists do *not* have to make costly *ontological** commitments such as those suggested by mentalists and rejected by behaviorists. An alternative appears once we recognize that (i) rational action consists in acting in accordance with the rules that characterize our practices and that (ii) access to these rules is secured by a specific kind of discursive practice: regulative folk-psychology. Second, it *can* be argued that 'rationality' and 'self-interest fulfillement' are not, in general, synonymous. Unconditional synonymy between the two terms is the outcome of a conflation: that between reasons for acting and the causes of an action. Once we observe that reasons for acting and the causes of an action are known to the actor in two distinct ways—the one practically, the other observationally—the conflation is easily dispelled. As contended by Sen, empirically relevant reasons for acting *need not* be restricted to the set of individually rational ones: an individual's reasons for acting may be moral, legal, individually rational, or collectively rational. Third, economists *can* benefit from philosophers' work. Psychology and biology, because they yield insights as to how behavior may fail to be rational, have already been accepted as relevant neighboring fields. Conceptual analysis and philosophy of mind, which can help unveiling the structure of our reasons for acting, have an equal claim to a place at the frontier of economics.[44] I concretize this third claim in Chapters 2 and 3.

[44]Harsanyi's analysis of rule-utilitarian voting behavior shows how the claim can be concretized (Harsanyi, 1977b). Sugden's and Bacharach's analysis of team reasoning (Sugden, 1993; Bacharach, 2006) and Roemer' analysis of Kantian reasoning (Roemer, 2010) equally are instances of this approach. By the same token, economists will maybe unveil new forms of irrational behavior, such as (i) when individuals act on their desires despite having, in point of fact, committed to conform their behavior to a different principle, or (ii) when individuals fail to realize that their action might not only be known under the description they approve of, but also under another description they do not approve of.

Chapter 2

On The Provision of Legitimate Public Goods

"When I perform my duties as a brother, a husband, or a *citizen* and carry out the commitments I have entered into, I fulfil obligations which are defined in law and custom and which are external to myself and my actions."

Émile Durkheim, The Rules of Sociological Method, 1895.

2.1 Introduction

According to the Oxford English Dictionary, a free rider is "a person who, or organization which, benefits (or seeks to benefit) in some way from the effort of others, without making a *similar* contribution." If "effort" refers to the magnitude of an individual's contribution and "similar" contributions are contributions of comparable magnitude, then there are instances in which revealed preference analysis of public good games does not appear to be empirically adequate.[1] Indeed, assume that revealed preference theorists have it right that rationality commands individuals to contribute in a way which maximizes their preferences, whatever these may be. Then, in the presence of rational individuals, (i) the realized distribution of contributions to the public good merely re-

[1] I refer here to the *a posteriori* version of revealed preferences. The revealed preference principle, in its "as if" or *a priori* sense, is not the least concerned with individuals' perception of the situation. Crucial aspects of the argumentation I put forth in this chapter will appear inappropriate to proponents of the latter approach. I discuss in Chapter 1 the (un)desirability and (un)tenability of the "as if" methodology. Arguments against it can equally be found in more authoritative sources. See, in particular, Sen (1993, 1997).

flects the realized distribution of individual preferences; (ii) no individual faces a *rational* obligation to contribute more than what she currently does; and (iii) there is no room for contributors to try to bring non-contributors to their senses. Yet, consider the following case: to finance state expenditures, a tax scheme has been instituted which, for feasibility reasons, is contingent on only a limited number of observables. Assume some individuals have a possibility to alter the status of their observable attributes *without actually undertaking the actions that theses observables are taken to be reliable symptoms of*.[2] For these individuals, a fraction of their taxes—to the extent that they do pay it—is, formally, a *voluntary* contribution to a public good. Why, then, is it not uncommon to witness, in such contexts, attempts by contributors to call non-contributors to their senses?

Can it be argued that the case just described is one in which individuals, despite having no *rational* obligation to increase their contribution, have a *moral* obligation to do so? In other words, neglecting the irrationality inherent in a failure to grasp the consequences of one's own moral standrads, are we in a case where contributors try to convince non-contributors that their preferences, as revealed, do not fulfill standards of morality they would freely abide by. Moral suasion happens, too, but it *not* the phenomenon I want to describe here. Indeed, the value of public expenditure is a complex matter over which we generally accept significant disagreements. Although many among us believe that *some* amount of public expenditure is needed for bringing about the common good; although moral arguments *sometimes* raise non-contributors' preferences for public expenditure and, by the same token, turn them into contributors; we *do* recognize that every rational non-contributor remains free *not* to conform her preferences to given moral standards.[3] On the contrary, we sometimes are reluctant to accept that she is under no obligation to limit tax avoidance. The reason is this: a tax scheme democratically arrived at reflects a *collective intention*, that is, a plan of action which citizens have agreed to *jointly* implement. Every non-contributor, *qua citizen*, has a right to mark her (rational) disagreement with the given plan but also an obligation to comply with it until it is democratically amended. The obligation is a *political* one[4] and is best seen, on Durkheim's words, as *external* to the individual.

[2]For instance, until January 1st 2016, a German citizen could, in order to lower her taxable income, legally engage in the dividend stripping practice known under the name of Cum-Cum trade (see, e.g., https://www.ft.com/content/741df8aa-178f-11e6-b197-a4af20d5575e.) The legality, prior to 2012 changes in German legislation, of a closely related practice, Cum-Ex trading, is still debated. Estimated costs of the two practices for European tax authorities add up to more than 50 billion Euros (https://cumex-files.com/en/).

[3]I leave aside the difficult question of the relationship between moral behavior and rational behavior. Some would argue that I cannot (Foot, 2001). But the question here is only meant to raise attention to a specific point: that complaints addressed by contributors to non-contributors, in the presence of *some* public goods, are the expression of a mode of reasoning that is not a moral one.

[4]Socrates did not base his rejection of Crito's offer to escape on moral grounds, but on political ones. See especially Plato (385), Sections 50c to 54e.

In this chapter, I suggest that, in a context of voluntary provision of a public good, such political obligations can fruitfully be formalized as steming from *a joint commitment to steer clear of free riding*. Despite my use of game theoretical tools, the analysis I present does not follow the tenet of revealed preferences. In fact, it cannot do so. Revealed preference theorists identify individuals' preferences with their choice behavior. In so doing, they *restrict* the interpretation of the game and the set of sensible solution concepts. If, say, in a two players game where Player 1 can take actions a_1 or b_1 and Player 2 actions a_2 or b_2, Player 1 is observed to choose (a_1, b_2) over (b_1, b_2), then, according to the principle of revealed preferences, *scientists can say no more than* that, in the event that player 2 opts for b_2, it is rational for player 1 to opt for a_1. In Chapter 1, I argue against this view on the ground that, *for any situation of interest, scientists have an opportunity to observe individuals' linguistic behavior too*. Pieces of linguistic behavior, for instance, may be an (approved of or disapproved of) application of the concept of free riding to qualify certain actions, or a (approved of or disapproved of) reference to a joint commitment. Call *legitimate* a public good such that, in its presence, individuals, with the help of these two concepts, bring their behavior in conformity with a system of political obligations. I claim that, in situations that involve a legitimate public good, preference revelation is *mediated* by our linguistic behavior in such a way that, to capture the situation's logic, we must depart from the revealed preference paradigm.

The distinction between legitimate and non-legitimate public goods is a concrete one. Consider the following two possible worlds: World 1 is populated by efficiency concerned individuals who may contribute to a public good by following a code of etiquette; World 2 is populated by altruists who may contribute to a public good by financing a charity that will help the least fortunate ones. Our own world appears to be a blend of these two worlds, but, importantly, *not any kind of blend.* For, when it comes to assessing whether behavior is in conformity with etiquette, individuals appeal to a *system of rights and obligations*, while they do not do so when they are given an opportunity to finance a charity. If, during a public lecture, I display good manners but you continually talk with your neighbor, then each of us would, in a normal context, aknowledge that I can authoritatively demand from you to 'behave.' Differently, my being a contributor to charities does not give me any such kind of entitlement: in a normal context, each of us would consider illegitimate any attempt from me to suggest that you have an *obligation* to contribute to charities. This discrepency, possibly, has to do with the fact that etiquette following behavior results from political considerations, while contribution to charities results from moral ones. At any rate, the presence of such systems of rights and obligations—call them *systems of demand-rights* (Gilbert, 2015)—regulating etiquette following behavior allows us to call efficiency a legitimate public good. The absence of such a

system for the case of charities, conversely, shows that we do not, in general, attribute a legitimate character to well funded charities.

Gilbert (1989, 1990, 2015) explains the presence of systems of demand-rights by appeal to a specific form of commitment: *joint commitments to act as a body*. Upon acting as a body, individuals no longer seek to *unilaterally* bring about their most preferred outcome, they commit to *multilaterally* carry out a plan of action. To model these situations, three approaches come to mind, according to whether the game form, the solution concept, or preference profiles are subjected to variations. Revealed preference theorists take as given the first two and look for individual preference patterns that generate empirically adequate outcomes. Insitutionalists take as given the two last and ask which game form yields sensible results. In this paper, I take preferences and game form as given and try to find a fitting solution concept. I develop a concept of collective equilibrium in which individual jointly commit to steer clear of free riding. The inherent vagueness associated with the "free rider" concept could, in principle, give rise to the presence of multiple equilibria. I show that, when systems of demand rights merely endow individuals with a right to remind lower contributors of their obligations, one equilibrium is more saliant than the others: that in which all individuals make identical contributions. I further argue that situations in which we might expect the appearance of such equilibria are situations in which participants form a relatively homogeneous group.

To sum up, I contend that departures from the revealed preference paradigm can increase our understanding of interactive situations. It is so, for instance, when the considered situation involves a specific kind of public good: a legitimate public good. In the presence of a legitimate public good, individuals jointly commit to steer clear of free riding. The idea that collective or moral commitments may play a significant role in public good provision was already pushed forward by several economists[5] and implemented by still others.[6] I implement it here in a novel way: none of the existing studies draws a distinction between legitimate and non-legitimate public goods, and, to the best of my knowledge, no model is outcome equivalent to the one I present here. In the next section I discuss the concept of legitimacy and briefly survey two literatures: that on the Nash theory of voluntarily provided public goods and that on collective intentionality and team reasoning. In section three, I present my main line of argument. Namely, that, *beyond* inefficiency, there is a *second* reason why individuals may want to put an effort to change their behavior or environment: *the presence of free riders*. I define free riding, investigate the determinants of free riding in Nash equilibrium, and present *collectively rational equilibria*, a solution concept which records the existence of a joint agreement to

[5]See, e.g., Runciman and Sen (1965), Sen (1973, 1974, 1977), Sugden (1982, 1993), or Bacharach (2006).
[6]See, e.g., Laffont (1975), Harsanyi (1982), Sugden (1984), or Roemer (2010, 2015).

steer clear from free riding. I show that the presence of a simple system of demand-rights may help relatively homogeneous individuals to reach an outcome that Pareto dominates the Nash outcome.

2.2 Legitimacy, Nash Reasoning, and Team Reasoning

2.2.1 Legitimate Public Goods

An unquestioned assumption permeates the literature on the voluntary provision of public goods; namely, that all public goods arouse identical forms of behavior. Given economic theorists' attachment to the principle of revealed preferences, this assumption need not come as a surprise; revealed preference theorists recognize the existence of *only one* form of rational behavior: internally consistent choice behavior. Yet, on the face of it, this assumption does run against a clear empirical observation: that specific systems of rights and duties are inherent to some situations of voluntary provision of a public good and not to others. To be concrete, assume you and I work in team on some project. Each of our efforts will increase the chances of a good outcome, and each of us cares about that. This is a case of voluntary provision of a public good, but not an arbitrary one. For, in such a situation, each of us can exert *demand-rights* to regulate the behavior of her or his teammate. If it becomes apparent that you are not in any way exerting efforts to bring the project forward while I am, I have the standing to remind you of your *engagements*. In normal circumstances, you will recognize that standing and make an effort to abide by your obligations. Matters differ when we consider the following alternative event. You and I are two regular consumers of the same youtube broadcast. The broadcaster makes a living through crowdfunding. I belong to the group of consumers who, every now and then, contribute a gift to this broadcaster and learn that you belong to the set of consumers who never contribute. I may feel uncomfortable about this. Nonetheless, in this event, I do not dispose of any demand-rights I could make use of to regulate your behavior.

Demand-rights, as defined in Gilbert (2015), are rights that a player has over "a particular action of a particular agent," and to have such a right is "to have a standing or authority to demand that action from the agent, and the standing or authority to issue appropriate rebukes to that agent should the action not be performed" (p. 23). To find an explanation for them, Gilbert analyzes familiar situations in which they are present. For instance, situtations of which we may say that two individuals are *walking together* (Gilbert, 1990). A strictly behavioral approach to the phenomenon of joint walks fails. A reason is this:

if, shortly after you set out to go for a walk, I *happen* to join you on the sidewalk, keep walking at a pace similiar to yours, and turn wherever you turn, after a while you will feel uncomfortable. This discomfort will arise because I behave *as if* we were walking together while, precisely, *we are not*. Thus, alternative accounts of walking together complement behavioral criteria with further necessary conditions. Typical among these are the requirements that each of the involved individuals *intends* to walk with the other one and that these intentions are *common knowledge*. Yet, these will not do either. For, Gilbert notes, it is characteristic of individual intentions that they may be unilaterally rescinded. So if it were the case that joint walks are sufficiently characterized by a set of commonly known individual intentions, I would not have any authority to complain upon your rescinding of your intention once we are in the middle of the woods. But I do. She concludes that the best explanation for the presence of demand-rights consists in taking them to be reliable symptoms of *joint commitment to act as a body*, a.k.a., a *joint intention*.

Accepting the idea of a joint intention demands some philosophical background.[7] I deal with this sensitive point in section 2.3 and, for now, come back to public goods. It is an interesting fact that some public goods (e.g., team work) arouse contributive behaviors regulated by systems of demand-rights while other public goods (e.g., youtube broadcasts) arouse supererogatory contributive behavior. Institutionalists may want to argue that modelers ought to integrate them in the game form. However, consider again the tax avoidance example I alluded to in the introduction. Assume the legislator, because she faces unreasonable enforcement costs, decides to grant identical sets of warranted actions to individuals with differing types. Had enforcement costs been sufficiently small, she would rather have incentivized the one type to act in this way and the other type to act in that way. In such a context, it is possible to draw a distinction between behavior that is *in line with the "spirit of the law,"* i.e., the set of states of affairs which the legislator attempts to bring about by means of a legislative act, and behavior *merely in line with the law*, i.e., the state of affairs in which each individual simply selects her favored option within her legally warranted set. In the first case, the actor uses her type to *infer* the set which, had enforcement costs been small enough, would have been legally waranted for her and selects her most preferred option within that set. She may be said to act *responsibly*. She does not do so in the second case, in which we may say of her that she acts *legally*. This distinction, I contend, casts doubt on the adequacy of an institutional approach.

There are notorious instances of legal but irresponsible actions, such as when individuals exploit the complex features of a tax (or quota) scheme to minimize tax payments (resp.

[7] Anscombe (1957), notably, refutes the identification of intentions with mental states. But her account involves a notion of practical knowledge. Can practical knowledge be plural?

maximize their profits). And I find interesting that, in media coverage of these instances, actors typically insist on the legal character of their deeds. Possibly, responsible behavior is frequently observed too. Consider the following statement found in Tirole (1999): "Almost every economist would agree that actual contracts are or appear quite incomplete.[8]" It is far from evident that contracting agents systematically exploit the incompleteness of their contracts. Nor, in the event that they havn't figured out the presence of 'free lunches,' that they would *want* to exploit them once someone points these out to them. Quite the reverse, consumers regularly contribute to a cleaner environment by diligently sorting their wastes or adapting their consumption behavior. In public hospitals, undersupplied medical staffs often do their best to maintain a high quality of service. In the private sector, employers screen out individuals unable to work in team and employees oftentimes carry out actions which are best described as being in the interest of their compagny, as opposed to their own. Explaining all such actions as irrational or as resulting from a congruence of preferences among the contracting parties need not be the best option. And introducing a concern for reputation only seems to postpone the issue.[9] It is possible, I suggest, to understand efforts to fulfil a contract's "spirit" as contributions to a legitimate public good.

2.2.2 Nash Equilibrium Theories of Voluntary Provision

A specific example of voluntary provision of a public good became prominent by raising difficulties for early Nash equilibrium theories. It is attributed to Kolm (1965), who argued that "the definition of the optimal distribution of welfare does not result from any value judgement made by the economist[, for he only] is an observer of citizens value judgements and opinions, as he is an observer of their tastes concerning consumers goods;" and that "the knowledge of these opinions presents exactly the same revelation difficulties as that of the tastes for public goods." In other words, if we start from the premiss that redistribution levels are a common argument in individuals' utility functions, then legal redistribution schemes constitute institutional provisions of a public good and charitable donations a volutary one. Following Sugden (1982), let me call a *public good theory of philantropy* any account of voluntary contributions to charities in which observed behavior is identified with the Nash outcome of a public good game. A central feature of these theories was shown to be deeply problematic: the perfect

[8]The reason, economists suggest, is that every situation that involves a contract, the law, or a custom, also involves transaction costs, and that these sometimes outweigh the benefits of contracting. Transaction costs arise from unforseen contingencies, design costs, or implementation costs.

[9]It is always possible to find instances where players act without being observed.

substituability, for every contributor, of gifts by others with her own gift.[10] Under two reasonable assumptions,[11] perfect substituability was shown (*i*) to rule out the possibility that large charities (e.g., the Red Cross) could exist (Sugden, 1982), (*ii*) to imply that each Euro of tax-financed governmental contribution to a public goods will crowd-out a Euro of private contributions (Warr, 1982; Bergstrom et al., 1986), and (*iii*) to imply that, in large populations of heterogeneously endowed individuals, the proportion of the population contributing to the public good decreases to zero: only the very richest contribute (Andreoni, 1988).

A simple, *ad hoc*, way of escaping the puzzle was formalized by Andreoni (1989, 1990). It consists in admitting that individuals are impure altruists, i.e., that it also matters to them which amount they *personally* contribute to the public good. Call "warm-glow" the influence that an actor's personal contribution has on her utility function, net of the impact it has through the induced increase in the amount of available public good. It is possible to remove the *ad hoc* character of warm-glow theory by showing that such a feature is a natural result of scientifically accepted features of human psychology. For instance, a concern for social reputation or self-image (Bénabou and Tirole, 2006, 2011). From this perspective, warm-glow theory brings about the desired outcome—imperfect substituability between own gift and gifts by others—in a respectable way. Yet, there are vantage points, too, from which warm-glaw theories are not entirely convincing. Elster (2011) argues that acting out of a concern for self-image involves a form of self-deception and that, as a consequence, warm-glow theories grounded in concerns for self image cannot form convincing theories of rational behavior. His argument is best conveyed by spelling-out a specific account of what it is, for an agent (person or group) to have an *intention*. On Anscombe's view (Anscombe, 1957), agent i intends to boost her self-esteem if and only if: (i) she tries to carry out action a, (ii) agent i belongs to a linguistic community in which action a can be taken to be a reliable symptom of "boosting one's self-esteem," and (iii) upon proceeding, i *practically knows*[12] action a under the description "boosting my self-esteem."

Now, assume agent i, by contributing a gift g_i to the public good, intends to boost her self-esteem. Then, Anscombe says, she must practically know her act of contributing a gift

[10] That is, if my utility function is measurable with respect to the quantities of public and private good I consume, then any gift to the public good by another player entails the same gross benefits as a gift to the public good I'd have made myself.

[11] That no contributor spends her entire income on the public good and that each of the public and private goods is, to every participant, a normal good.

[12] It is a central feature of Anscombe's view that an agent *practically knows* what she intends to do. In a nutshell, this say that, upon acting intentionally, we do not know the state of affairs we try to bring about through observation but, rather, we use an existing description of state of affairs, e.g. 'boosting our self-esteem,' as a blueprint for our action. This form of knowledge gives an impression of *groundlessness*, unlike when we or others get to know something as observers.

under the description "boosting my self-esteem." On the other hand, if her intention is successfully carried out, she must also be taking her action to be a reliable symptom of her moral skills. That is, she must *observationally know* her contribution under a description of the form "a (morally) good action." Self-deception, then, consists in the following fact: agent i successfully intends to increase her self-esteem by contributing to a charity if and only if the description under which she practically knows her contribution is in conflict with the description under which she observationally knows it. Although no such problem arises with warm-glow based on reputational concerns—reputational warm-glow will occur whenever a similar conflict arises between agent i's practical knowledge of her actions and *other agents*' observational knowledge of them—the issue is worth attention: not all situations are such that the agent can, by her deeds, affect her reputation. In fact, Andreoni himself is not reluctant to associate warm-glow to an emotional, as opposed to intentional, attitude. This is illustrated by his statements (Andreoni et al. (2017), p. 627):

> Psychologists posit that giving is initiated by a stimulus that elevates sympathy or empathy in the mind of the potential giver, much as the smell of freshly baked bread can pique appetite. Resolving this feeling comes either by giving and feeling good or by not giving and feeling guilt.

I have argued in the introduction that cases of voluntary contributions to a charity are different from cases of voluntary contributions to a legitimate public good. The claim that, in the context of donations to charities, individuals act on their emotionans is credible and has gained empirical support (Ribar and Wilhelm, 2002; Crumpler and Grossman, 2008; Ferguson et al., 2012). Nonetheless, this does not give a reason why *every* instance of voluntary contributions to a public good should be best associated with warm glow. In particular, it remains important to distinguish motives from the byproducts of an action.[13] In the sequel, I want to suggest that, in normal cirumstances, an individual's motive for contributing to a *legitimate* public good should be looked for in her sense for collective rationality rather than in her emotions.

[13]From the general presence of positive and negative feelings in situations that invite individuals to make a donation, one cannot infer that the motive with which one acts reduces to search or avoidance of such feelings. When reacting to epicurean philosophies, Seneca (see Seneca (8 AD), Book IX) found a poetical way to express this point: " —'But,' says our adversary, 'you yourself only practise virtue because you hope to obtain some pleasure from it.' —In the first place, even though virtue may afford us pleasure, still we do not seek after her on that account: for she does not bestow this, but bestows this *to boot*, nor is this the end for which she labours, but her labour wins this *also*, although it be directed to another end. As in a tilled-field, when ploughed for corn, some flowers are found amongst it, and yet, though these posies may charm the eye, all this labour was not spent in order to produce them–the man who sowed the field had another object in view, he gained this over and above it–so pleasure is not the reward or the cause of virtue, but comes in addition to it; nor do we choose virtue because she gives us pleasure, but she gives us pleasure also if we choose her." (Also quoted in Elster (2011), emphasis added.)

2.2.3 Collective Intentions and Team Reasoning

Understanding which motives *actually* guide contributions to a public good, Sugden argued, may demand from us to "drop the assumption of utility maximization" (see Sugden (1982), p. 349). There are, to the best of my knowledge, two economic theories of public good provision that do so: one is Sugden's theory of moral reciprocity (Sugden, 1984); another is Roemer's (normative) theory of Kantian behavior (Roemer, 2010, 2015). The theory I present here constitutes a third alternative. I draw a distinction between legitimate and non-legitimate public goods and suggest that, in the presence of a legitimate public good, contributing individuals jointly commit to bring about an outcome free from free riding. Call such a joint commitment a *collective intention to steer clear of free riding*. Collective intentions are reminiscent of a distinction that was put forward by Runciman and Sen (1965) in the analysis of prisonners' dilemmas: Rousseau's distinction between the "general will" and the "will of all" (Rousseau, 1755, 1762).[14] I do identify collective intentions to steer clear of free riding with outcomes of a specific kind of commitment by a general will. Namely, a commitment to bring about a minimal form of common good: a public good provision process exempt of free riding. The notion of collective intentions raises both scientific and philosophical concerns; I clarify these in this section.

To start with scientific concerns; one may ask how, empirically, situations involving a collective intention are distinct from situations involving a collection of individual intentions. Such concerns can be met, for differences occur along two dimensions: a conceptual one and an empirical one. To start with the former, call *equilibrium path* any collection of mutually compatible contingent plans of actions and reasons for these plans. It is clear that reasons that allude to a general will *need not* bring support to the same equilibrium paths as those to which reasons that exclusively appeal to individual wills bring support. For instance, if we take Pareto efficiency to be a necessary condition for identifying an outcome with "the common good,[15]" then the exertion of the general will yields an outcome on the Pareto frontier. This need not be true of the will of all which, as long as individuals abide by Bayesian rationality, is bound to yield a correlated equilibrium (Aumann, 1987).[16] The second difference is empirical: reference to a general will allows

[14]Given a collection of individuals, the general will finds expression in a collective intention to bring about the common good while the will of all amounts to every individual intending to bring about her preferred state of affairs. Runciman and Sen suggest that, in a prisonners' dilemma, a "conflict arises between the will of all [all confess], they note, and the general will [noone confesses] [...] because of a difference between the outcome of individual strategy and of enforced collusion."

[15]In the sequel, I will not do so.

[16]This is precisely the point that Runciman and Sen made for the case of a prisonner's dilemma. Followers of Samuelson, eventually, will find this difference irrelevant. Indeed, for them, *any* meaningful statement may constitute a reason for acting. If one accepts this contention, it does seem that every path

to parsimoniously explain the the *authority* that individuals have when exerting of systems of demand-rights *out of the equilibrium path*. In the absence of a collective intention there is no room, out of the equilibrium path, for one player to *authoritatively* demand from another player that she changes her behavior. At best, one player will have the standing to *suggest* another player that her action may not be in line with her preferences. But the concerned player retains the authority to deny or confirm that suggestion. In the presence of a collective intentions systems of demand-rights naturally arise because, contrary to individual intentions, collective intentions may *not* be unilaterally rescinded. The jointly committed players have an obligation to bring their actions in conformity with the others' normative expectations.

The second set of concerns is philosophical: it is not clear that the concept of intentions, which, at first sight, involves mental states, can be applied to groups of individuals without committing the user to a demanding ontological position. More precisely, an answer must be given as regard to (i) the identity of the agent to which the collective intention is attached and (ii) the ontological status of that subject. These concerns have been partially answered in the literature. Rousseau suggests that the general will is the will of a civic body, a "public person [...] formed by the union of all persons" (Rousseau (1762), Book I, Chapter vi). But one does not have to look for remote philosophers. In the area of modern analytical philosophy, Gilbert (1989, 1990) builds an anchor point for Rousseau's thought. She argues that the general will is the will of a "plural subject" that cooperators have jointly agreed to constitute. And, importantly, her conclusion does not amount to a claim about the *actual* existence of plural subjects. The reason is twofold. First, in expressions of intentions, the pronouns 'I' or 'we' need not have a referring use (Anscombe, 1975). One may, for instance, rather think of them as indexes for the involved form of reasoning: individual or team. It is shown in Gold and Sugden (2007) that, provided sufficient common knowledge conditions are fulfilled, there are team reasoning schemata that can be expressed from the viewpoint of an individual team member. Second—and this may be of help to circumvent demanding common knowledge assumptions—the layman's identification of intentions with causal mental states, Anscombe (1957) showed, is flawed. Intentions conceived along Anscombian lines, that is, intentions in action,[17] may be fit for a use in the plural mode (Schmid, 2016, 2018).

singled out by collective reasons can equally be singled out by *some* collection of individual reasons. I argue in Chapter 1 that *it is not the case that* any *meaningful statement can constitute a reason for acting*.

[17] See Chapter 1.

2.3 The Model

2.3.1 Homogeneous, Linear Public Good Economies

An 'economy,' in the everyday use of the word, can be thought of as entailing a set I of n individuals, a set $W \equiv \times_{i\in I} W_i$ of individual endowments in inputs (a.k.a., resources) a set X of conceivable outputs (a.k.a., goods), and, for each individual i in I, an individual technology correspondence $f_i(\cdot)$ and a preference function $u_i(\cdot)$. The former maps input quantities into a subset of the output space, $f_i(W_i)$, called individual i's feasible output set. The latter maps output vectors into $\mathbb{R}$, individual i's utility space. I the sequel, I restrict the meaning of the word 'economy' to these very kinds of tuples, $e \equiv \langle I, W, X, (f_i(\cdot))_{i\in I}, (u_i(\cdot))_{i\in I}\rangle$, and denote $\mathcal{E}$ the universe that contains all such possible economies. I restrict my attention to a specific subset of this universe, namely, that of well-behaved, homogenous and linear, public good economies, $\mathcal{E}^{PG}$.

Definition 2.1. (*Well Behaved, Homogeneous, Linear Public Good Economies*)

An economy $e \equiv \langle I, W, X, (f_i(\cdot))_{i\in I}, (u_i(\cdot))_{i\in I}\rangle$ *is a member of* $\mathcal{E}^{PG}$ *if and only if:*

(*i*) $W \equiv [0, \bar{w}]^n$ *for some* $\bar{w}$ *in* $\mathbb{R}_{++}$*;*

(*ii*) $X \equiv \mathbb{R}^{n+1}_+$ *and has typical element* $(x_0, x_1, \cdots, x_n)$*;*

(*iii*) $f_i(\cdot)$ *maps any element* w_i *of* $[0, \bar{w}]$ *into the subset*

$$f_i(w_i) \equiv \left\{(x_{0,i}, x_i) \in \mathbb{R}^2_+ \mid x_i + c_i x_{0,i} = w_i\right\},$$

where c_i *is an element of* $(1, +\infty)$*.*

(*iv*) *Given* $x_0 \equiv \sum_{i\in I} x_{0,i}$*, there is an increasing, twice differentiable function* $u(\cdot)$ *from* $\mathbb{R}^2_+$ *to* $\mathbb{R}$ *such that, for all* i *in* I *and every* x *in* $\mathbb{R}^{n+1}_+$*,*

$$u_i(x) \equiv u(x_0, x_i).$$

Furthermore, $u(\cdot)$ *is strictly concave in each argument and for all* (x_0, x_i) *in* $\mathbb{R}^2_+$*,*

$$\frac{\partial^2 u(x_0,x_i)}{\partial x_0 \partial x_i} \geq 0$$

Assumption (iv) captures the fact that x_0 is a public good and that the x_is are private goods. The measurability restriction on $u(\cdot)$ rules out explicit concerns of an individual for others' private consumption levels. The second order condition focuses the analysis on situations in which the public and private commodities are *not* substitutes. Two further restrictions worth noting are those on individual endowments and utility functions:

I assume they are homogeneous. There does remain a source of heterogeneity; namely, I allow idiosyncracies in individuals' productivites in providing the public good. Without loss of generality, let individual indexes be such that $c_1 \leq c_2 \leq \cdots \leq c_n$.

For all w in W, $f(w) \equiv \times_{i \in I} f_i(w_i)$ and $x_0 \equiv \sum_{i \in I} x_{0,i}$ jointly characterize the set of feasible public and private good allocations feasible upon feeding in w to the production process. Note that well-behaved public good economies allow for both *wasteful* allocations, i.e., alternatives in $X \setminus f(\bar{w})$, and *expropriative* allocations, i.e., alternatives in which some of the x_i's, i different from 0, are null. Whether or not such allocations are picked out depends on the set of *formal institutional rules* that hold in economy e. Call any collection of (i) a number of players, (ii) action sets (one per player), and (iii) preference functions (one per player) a *game*. Conventionally, formal institutional rules are explicitely modelled by the means of a *game form*: a function that maps the set of possible economies into a set of possible games. I focus here on game forms which associate to every well behaved public good economy what is commonly called a voluntary public good game.

Definition 2.2. (*Voluntary Public Good Games*)

For every economy e in $\mathcal{E}^{PG}$, a game form $G(\cdot)$ is said to define a voluntary public good game if and only if $G(e) \equiv \langle I, (A_i)_{i \in I}, (u_i)_{i \in I} \rangle$ where, for all i in I,

(i) $A_i \equiv \{(x_{0,i}, x_i) \in [0, \bar{w}]^2 \mid x_i + c_i x_{0,i} = \bar{w}\}$, *and*

(ii) $x_0\left((x_{0,i})_{i \in I}\right) = \sum_{i \in I} x_{0,i}$ *and, for all i in I and a in $A \equiv \times_i A_i$*

$$u_i(a) = u(x_0, x_i).$$

I denote $\mathcal{G}^{PG} \equiv \left\{G(e) : e \in \mathcal{E}^{PG}\right\}$ the collection of all conceivable public good games in well behaved, homoegeneous, linear public good economies.

In words, in a public good game, each individual i has full control over her endowment $\bar{w}$.[18] She may use it to increase her private consumption x_i or in order to make a gift $x_{0,i}$ to the production of the public good x_0. The actual cost of a gift $x_{0,i}$ to individual i, in turn, faithfully reflects the production technology available to the individual.

2.3.2 Free-Riding: Nash Equilibrium Comparative Statics

I now consider an arbitrary economy e in $\mathcal{E}^{PG}$ in which formal institutional rules define a public good game $G(e)$. A well-established assumption, in economics, is that neither

[18] The requirement that each individual budget constraint binds is without loss of generality in well behaved public good economies.

the specific aspects of the economy (e.g., the presence of a public good or not, its being a substitute or a complement to private goods, etc.) nor the details of formally established institutional rules (e.g., whether they define a public good game, whether there is a possibility to exclude some individuals from consumption, etc.) are relevant to the determination of a rational individual's rule of behavior in the game. Economists start from the premise that rational individuals necessarily (or eventually) abide by Nash's behavioral postulates or, at any rate, that rationality commands them to do so. That is, each player takes others' behavior as given and, in response to it, selects the element in her action set which maximizes her individual preferences. In a Nash equilibrium of a public good game, therefore, each individual virtually solves a standard consumption problem with an alterred non-negativity constraint (Sugden, 1982; Bergstrom et al., 1986). Letting $x_0^{-i,*} \equiv \sum_{j \neq i} x_{0,j}^*$ denote the equilibrium provision of public good by players other than i, the problem may be formally stated as follows:

$$(\mathcal{P}_i^{\text{NE}}) \qquad \begin{aligned} &\underset{(x_0, x_i)}{\text{maximize}} && u_i(x_0, x_i) \\ &\text{subject to} && x_i + c_i x_0 = \bar{w} + c_i x_0^{-i,*} \\ & && x_0 \geq x_0^{-i,*}, \; x_i \geq 0 \end{aligned}$$

A Nash equilibrium is an allocation x^* in X such that, for every individual i in I, (x_0^*, x_i^*) solves $\mathcal{P}_i^{\text{NE}}$. Under the prevailing assumptions, a Nash equilibrium can be shown to exist and be unique (Bergstrom et al., 1986). Following conventional uses of the term "free rider," as recorded in the Oxford Dictionaries, I suggest to formally define free riding as follows:

Definition 2.3. (Free Rider)

For any given tuple $(e, G(e))$ in $\mathcal{E}^{PG} \times \mathcal{G}^{PG}$, we may fix a λ_e in $\mathbb{R}_+$ such that, in economy e and game $G(e)$, individual i is called a free rider *at allocation x if and only if*

$$x_{0,i} \equiv \tfrac{\bar{w} - x_i}{c_i} < \tfrac{\lambda_e}{n} x_0,$$

λ_e is called the standard *of economy e in game $G(e)$.*

λ_e may be seen as marking out an informal, linguistic convention which prevails in economy e and game $G(e)$. $\lambda_e = 1$ corresponds to a case where all individuals with a contribution lower than the average contribution are judged to be free riders. $\lambda_e = 0$ corresponds to a case where the free riding concept has no application. In numerous situations, individuals seem to accept some degree of variation in objectively measured individual contributions. This suggests a standard λ_e strictly smaller than unity. Con-

versely, situations can be observed, too, in which individuals' ability to accept choices that differ from theirs shows limits. In such situations, the standard λ_e lies strictly above it's lower bond, 0. More generally, in the presence of a legitimate public good, if an individual's contribution is, for no observable reason, significantly lower than the average contribution, a linguistic convention exists which entitles agents to call that individual a free rider.

Now, concepts are devised and applied for a reason: they enable members of a linguistic community to evaluate a situation and, if judged necessary, to consider alternative institutional arrangements. If, in the presence of a public good, the occurence of free-riding is taken to be a reliable signal of a deficient situation, one may expect that the setting up of formal or informal institutional arrangements will depend on the number of free riders in equilibrium. In this sense, a characterization of the situations in which we may expect participants to come to a conclusion of deficiency, *on the free-riding dimension*, would prove most useful. A preliminary observation suggests that heterogeneity in realized cost types matters. Indeed, if there exists a c in $(1, +\infty)$ such that, for all i in I, c_i equals c, then all equilibrium contributions are identical, that is, there are no free riders. Therefore, I define:

Definition 2.4. (Cost Homogenization)

Consider an economy e in $\mathcal{E}^{PG}$ and let c denote the realized vector of cost types in e. Let α be an element of $[0, 1]$. I call α-homogenization *of c the vector of cost types $\tilde{c}$ such that $\tilde{c}_1 = c_1$ and, for all i in $I \setminus \{1\}$,*

$$\tilde{c}_i - \tilde{c}_{i-1} = \alpha(c_i - c_{i-1}).$$

I call α-homogenization *of e the economy $\tilde{e}$ with cost type realization $\tilde{c}$ that is otherwise identically equal to e.*

General results on comparative statics prove hard to come by. But the following one obtains:

Proposition 2.1. *Let $\mathcal{E}^{CD}$ denote the subset of $\mathcal{E}^{PG}$ such that, for all e in $\mathcal{E}^{PG}$, individual preferences can be represented by a Cobb-Douglas utility function. Assume there is a λ in $(0, 1]$ such that, for every economy e in $\mathcal{E}^{CD}$, $\lambda_e = \lambda$. Starting from an economy e in $\mathcal{E}^{CD}$, the equilibrium number of free riders remains unchanged or decreases whenever:*

(*i*) *$\bar{w}$ decreases;*

(*ii*) *c homogeneously increases; or*

(*iii*) *for α small enough, c is α-homogenized.*

2.3.3 Systems of Demand-Rights and Joint Intentions

It is well known that, in the presence of a public good, Nash behavior is likely to bring about an *inefficient* outcome. Inefficiences, when they come to be recognized, constitute one reason why individuals think about amending the environment they face. The point of the previous section was to show the following: there are situations such that, if every individual abides by a personal commitment to bring about her most preferred outcome, then, mechanically, *a significant fraction of the players will be free riders.* As I see it, this constitutes an *additional reason* why individuals may wish to bring about policy or cultural changes. Indeed, inefficiencies and the presence of free riding are *distinct* phenomena. This may be seen by noting that Nash equilibria of public good games without idiosyncracies are instances in which the first occurs but not the second. And that an instance of the converse arises if I provide the Pareto optimal amount of public good while you twiddle your thumbs. In the presence of free riding, therefore, we may expect individuals to think about amending their environment. On the one hand, they could adjust the rules of the game, that is, set-up a system of formal incentives. This is the kind of procedures investigated in the instituional literature (Ostrom, 1990); I shall not delve into these here. On the other hand, they could opt for *jointly committing to act as a body*. In this essay, I am concerned with this second kind of procedures.

It is uncontroversial that individuals regularly impose constraints on their own behavior. We call these *personal commitments* (*of the will*) or, more commonly, *intentions*. Joint commitments are best understood when contrasted with personal commitments. For, in a similar manner, "two or more people [can] impose [a] commitment on the same two or more people– *as one*" (Gilbert (2015), p. 21). Call a joint commitment to act as a body a *joint intention*. Two points are worth noting about them. First, it can be argued that the use by individuals of a system of mutual demand-rights *is* a reliable symptom of the presence of a joint intention (Gilbert, 1990). The reason is that joint intentions, unlike personal ones, are *mutually* agreed upon. As a consequence, they may only be *multilateraly* legitimately rescinded. So, in any given situation, a unilateral deviation from a joint intention by one individual entitles conforming individuals to make use of demand-rights and regulate her behavior. Second, when individuals jointly intend to bring about some state of affairs, they no longer seek to conform their behavior to a principle of individual rationality; rather, each is committed to *do her bit* in the collective action in which she has engaged. This, eventually, raises a question: what does it *mean*

for someone to be doing her bit? The answer will depend of the common goal that the individuals set to themselves. I contend that, in the presence of a legitimate public good, individuals *jointly intend to steer clear of free riding*.

To be precise, I contend that individuals, rather than committing to bringing about a *specific* amount of public good provision or to bringing about a Pareto efficient outcome, commit to adjust their behavior whenever the latter singles them out as free riders. Consider a well defined public good economy e in $\mathcal{E}^{PG}$ in which formal institutional rules define a public good game, $G(e)$.

Definition 2.5. (*System of mutual demand rights*)

A system of mutual demand-rights $\lambda(\cdot)$ *is in place in economy* e *if and only if there exists a map* $\lambda(\cdot)$ *from* $(1, +\infty)$ *in* $\mathbb{R}_+$ *such that, for all* i *and* j *in* I *with* i *different from* j*, if* $x_0^i \geq \frac{\lambda(c_i)}{n} x_0$ *and* $x_0^j < \frac{\lambda(c_j)}{n} x_0$*, then* i *has a right to demand from* j *that she increases her effort,* j *recognizes the legitimacy of this demand and corrects her behavior in such a way that she can no longer be called a free rider.*

If a realized allocation involves free riding, then implementing a demand-right *once* will bring about an allocation in which one person among the free riders has increased her contribution in such a way that, given other's contributions, she may no longer be called a free rider. If that person wasn't the only free rider, then the system of demand rights still has application. But even in the event that this person was the only free rider, the system of demand right still may have application. Indeed, the increase in her contribution, by raising the average, may affect the free riding status of individuals who, this far, wern't free riders. In fact, if there is a second free rider correcting her behavior, then the first free rider's adjustment will not be sufficient to maintain her 'non-free riding' status. She will, once more, be subject to an obligation to correct her contribution. I call the generated process an *exhaustive implementation of the system of mutual demand-rights.*

Note that, starting from an allocation x in X, the presence of a system of demand-rights $\lambda(\cdot)$ induces a unique partition of the set I of players into two subsets, I_λ and $\bar{I}_\lambda$, such that individual i is in I_λ if and only if an exhaustive implementation of individuals' demand-rights would eventually oblige her, at some point, to alter her contribution. Call every such individual an *eventual free rider*. It is natural to assume that, when judging whether individual i's move counts as a bit or not, what actually matters isn't the starting allocation but whether or not she is an eventual free rider.

Definition 2.6. (Doing one's bit)

For any given tuple $(e, G(e))$ *in* $\mathcal{E}^{PG} \times \mathcal{G}^{PG}$*, individual* i *is said to be* doing her bit *at allocation* x *in* X, in the presence of a system of mutual demand-rights $\lambda(\cdot)$ *if and only if she isn't an eventual free rider.*

2.4 Collective Equilibria

A legitimate exertion of a demand-right, eventually, brings about a change in the allocation that opened an opportunity for this exertion. It is natural, therefore, to identify such events to equilibrium off-path behavior. Since (expectations about) off-path behavior shapes on-path behavior, we may ask how equilibrium on-path behavior looks like in the presence of a system of mutual demand-rights. I focus on a simple case: that of cost-independent systems of demand rights. I show that, in this case, a specific form of equilibrium behavior is salient; namely, one in which all players contribute identical amounts to the public good. I further show that such systems of demand-rights are most appropriate when the degree of heterogeneity across individuals remains limited.

2.4.1 Constant Collective Equilibria

Consider the case of a system of demand-rights $\lambda(\cdot)$ that is independent of individuals' cost types realizations. That is, assume there exists a λ in $[0, 1]$ such that, for all i in I, $\lambda(c_i) = \lambda$. I shall refer to this system as the *constant system of mutual demand-rights* λ. In the presence of a constant system of mutual demand-rights, doing one's bit can be characterized as follows:

Proposition 2.2. *For any given tuple* $(e, G(e))$ *in* $\mathcal{E}^{PG} \times \mathcal{G}^{PG}$*, individual* i *is* doing her bit *at allocation* x *in* X in the presence of a constant system of mutual demand-rights λ *if and only if*

$$x_{0,i} \geq \frac{\lambda}{n - |I^i_-|\lambda} \left(\sum_{j \in I^i_+} x_{0,j} \right) \qquad (\star)$$

where $I^i_- \equiv \{i\} \cup \{j \in I : x_{0,j} < x_{0,i}\}$ *and* $I^i_+ \equiv \{j \in I : j \neq i, x_{0,j} \geq x_{0,i}\}$.

To investigate stability considerations, it is helpful to consider a two stage process that replicates the dynamics entailed by the exhaustive implementation of the constant system of demand-rights. Assume that, in stage 1, each individual i in I can choose a quantity $x_{0,i}$ to contribute to the public good and that, in stage 2, every individual has to *consent* or (legitimately) *complain*. Also, assume that in the event that an individual has legitimately

complained, the system of mutual demand-rights is put to work and all eventual free riders are called upon to adjust their contributions. Let x denote the allocation reached by the end of stage one and denote

$$\bar{k} \equiv \arg\max_{k \in I_\lambda} x_{0,k},$$

the index of the highest contributor among eventual free riders. In the event of a second round complaint, a new allocation $\tilde{x}$ will be reached such that:

$$\tilde{x}_{0,i} = \begin{cases} \frac{\lambda}{n - |I_-^{\bar{k}}|\lambda} \left(\sum_{j \in I_+^{\bar{k}}} x_{0,j} \right) & \text{if } i \in I_\lambda \\ x_{0,i} & \text{else} \end{cases}$$

For any x in X, call $\tilde{x}$ the *legitimate adjustment* of x. A *Constant Collective Equilibrium* (*CCE*) differs from a Nash equilibrium in essentially one aspect: individuals base their considerations about whether or not they should deviate on *different counterfactuals*. More precisely, when considering whether or not she has a reason to deviate, a rational individual does not, as a Nash reasoner would, compare the prevailing allocation to the one she would be bringing about by unilaterally deviating; rather, she compares the prevailing allocation to *the legitimate adjustment* of the allocation she would be bringing about by unilaterally deviating.[19] This alternative counterfactual isn't a piece of magical thinking, but the necessary consequence individuals' common knowledge of the presence of a joint commitment to steer clear of free riding. In other aspects, the equilibrium concepts are similar. In particular, an allocation marked out as 'equilibrium' is one in which no individual has a reason to deviate. More formally:

Definition 2.7. (CCE – Constant Collective Equilibrium)

Fix a tuple $(e, G(e))$ *in* $\mathcal{E}^{PG} \times \mathcal{G}^{PG}$. *A* Constant Collective Equilibrium (CCE) *is an allocation and constant system of mutual demand-rights pair* (x^*, λ) *in* $X \times \mathbb{R}_+$, *such that, for all* i *in* I,

(i) i *isn't an eventual free-rider, i.e.,*

$$x^*_{0,i} \geq \frac{\lambda}{n - |I_-^i|\lambda} \left(\sum_{j \in I_+^i} x^*_{0,j} \right) \qquad (\star)$$

(ii) i *does not have a reason to deviate, i.e., for all* $x_{0,i}$ *in* $[0, \bar{w}]$,

[19] I am not the first one to suggest that Nash behavior *need* not constitute an adequate lense to understand social behavior in the presence of a public good (see, in particular, Sugden (1982, 1984) and Roemer (2010, 2015)). The solution concept I present here differ from theirs. Subsists the question whether or not a *scientific* investigation can make use of alternative lenses. On that matter, I point out in Chapter 1 some flaws in what I take to be two important members of the set of rationales which entice economists to understand Nash's behavioral postulates along *as if* lines. For more authoritative criticisms of *as if* interpretations of game theory, see e.g., Sen (1993, 1997) or Hausman (2011). For evidence that, in actual practice, the *as if* interpretation does *not* prevail see, e.g., Coase (1982), McCloskey (1983), and Dietrich and List (2016).

$$u\left(x^*_{0,i} + \sum_{j\neq i} x^*_{0,j}, \bar{w} - c_i x^*_{0,i}\right) \geq u\left(\sum_{j\in I} \tilde{x}_{0,j}, \bar{w} - c_i \tilde{x}_{0,i}\right)$$

where $\tilde{x} \equiv (\tilde{x}_{0,i})_{i\in I}$ *is the legitimate adjustment of* $(x_{0,i}, x^*_{0,-i})$.

I now show, for the case of two players, that a saliant CCE exists in which all individuals are required to contribute the same amount and the quantity provided corresponds to the one that maximizes the lowest cost individual's preferences. Consider the repeated game mentionned above. In stage 1, each of the two players decides how to allocate her income between public and private consumption. In stage 2, each of the two players, after observing the vector of first satge contributions, chooses between giving her consent (ct) and complaining (ca). In the event that all players consent, the game stops. In the event that one player complains, the legitimate adjustment of the first stage outcome is brought about before the game stops: every individual who, according to first stage contributions, is an eventual free rider, adjusts her contribution so as to meet her obligations. Once the game has stopped, payoffs are realized. Formally, for each $i \in \{1, 2\}$, let

$$A^1_i \equiv \{x_{0,i} \mid x_{0,i} \in \mathbb{R}_+\}$$
$$A^2_i \equiv \{ct, ca\}$$

respectively characterize the first and second stage action spaces of each player. For each i in $\{1, 2\}$, an element $x_{0,i}$ of A^1_i represents the amount of public good that individual i offers to finance in the first stage.

I proceed by backward induction, that is, I first consider the second stage. At this point, the realized vector of first stage contributions, $x_0 = (x_{0,1}, x_{0,2})$, is observed and, given some λ in $[0, 1]$, both players are indifferent between complaining and consenting if and only if

$$\mid x_{0,i} - x_{0,j} \mid \leq \left(1 - \frac{\lambda}{2-\lambda}\right) \max\{x_{0,i}, x_{0,j}\} \tag{2.1}$$

Indeed, only in such instances does x_0 coincide with its legitimate adujstment. In other instances, the player contributing the highest amount, say, player i, has an individually rational incentive to complain; for, whenever

$$x_{0,i} \geq \tfrac{\lambda}{2-\lambda} x_{0,i} > x_{0,j}$$

the legitimate adjustment of x_0, $\tilde{x}_0 = (\tilde{x}_{0,1}, \tilde{x}_{0,2})$ is such that $\tilde{x}_{0,i} = x_{0,i}$ and $\tilde{x}_{0,j} > x_{0,j}$. We may now turn to the first stage and keep in mind that, in any equilibrium involving consent by all players, the pair (x, λ) satisfies equation (2.1).

In the first stage, given any contribution $x_{0,j}$ by player j, we can view player i as choosing between two options:

(i) Play an adjusted Nash best response $\tilde{BR}_i(x_{0,j}) \equiv max\{BR_i(x_{0,j}), \frac{\lambda}{2-\lambda}x_{0,j}\}$ to $x_{0,j}$. That is, Nash best respond as long as this does not lead to missing on one's obligations and, otherwise, merely abide by one's obligations. This first stage strategy yields her:

$$u_i^{\tilde{BR}}(x_{0,j}) = u\left(x_{0,j} + BR_i(x_{0,j}), w - c_i BR_i(x_{0,j})\right) =: u^{BR}(x_{0,j})$$

as long as $BR_i(x_{0,j}) \geq \frac{\lambda}{2-\lambda}x_{0,j}$ and

$$u_i^{\tilde{BR}}(x_{0,j}) = u\left(\tfrac{2}{2-\lambda}x_{0,j}, w - c_i\tfrac{\lambda}{2-\lambda}x_{0,j}\right) =: u^{A}(x_{0,j})$$

otherwise.

or,

(ii) Take the lead, that is, pick a contribution $\bar{x}_{0,j}$ greater or equal to $x_{0,j}$ for player j, contribute $\frac{2-\lambda}{\lambda}\bar{x}_{0,j}$, and complain in the second round to have player j abide by $\bar{x}_{0,j}$. This first stage strategy yields her:

$$u_i^{L}(x_{0,j}) = \max_{\bar{x}_{0,j} \geq x_{0,j}} u\left(\tfrac{2}{\lambda}\bar{x}_{0,j}, w - c_i\tfrac{2-\lambda}{\lambda}\bar{x}_{0,j}\right)$$

The next lemma states that it is individually rational for player i to base her choice between the two strategies on a threshold.

Lemma 2.1. *For any given λ in $(0, 1]$, there exists a unique threshold $\hat{x}^i_{0,j}(\lambda)$ in $\mathbb{R}_+$ such that player i strictly prefers taking the lead over playing her adjusted Nash best response if and only if player j's contribution, $x_{0,j}$, is smaller than $\hat{x}^i_{0,j}(\lambda)$.*

Denote $\bar{x}^{i,L}_{0,j}$ the unconstrained maximizer of the lead utility function.[20] Intuition may be gained by considering two facts. First, since best response contributions are decreasing in the opponent's gift, the relative *cost* of contributing $\frac{2-\lambda}{\lambda}\bar{x}^{i,L}_{0,j}$, as opposed to one's best response, is *increasing in the opponent's gift*. This obtains because $\bar{x}^{i,L}_{0,j}$ is independent of $x_{0,j}$, so that the difference between $\frac{2-\lambda}{\lambda}\bar{x}^{i,L}_{0,j}$ and i's Nash best response increases with $x_{0,j}$.[21] Second, since selecting $\frac{2-\lambda}{\lambda}\bar{x}^{i,L}_{0,j}$ over the Nash best response brings about a constant contribution level on the part of j, $\bar{x}^{i,L}_{0,j}$, the relative *benefit* of picking it, as opposed to the Nash best response, is weakly *decreasing in the opponent's gift*. Thus, when player i's opponent contributes little, player i's best response is at it's closest from $\frac{2-\lambda}{\lambda}\bar{x}^{i,L}_{0,j}$; she

[20] That is, disregarding the fact that, if i is to take the lead, $\bar{x}^{i,L}_{0,j}$ must be greater or equal to $x_{0,j}$.

[21] If $x_{0,j} \geq \bar{x}^{i,L}_{0,j}$, the contraint binds and the maximizer simply is $x_{0,j}$ itself. The identity map being increasing, the logic still applies.

is ready to pay the (relatively) small cost involved in picking it over her best response because she gains a (relatively) large benefit from doing so. As the opponent's contribution increases, the gain she forgoes by not sticking to her best response increases and the benefit from selecting $\frac{2-\lambda}{\lambda}\bar{x}_{0,j}^{i,L}$ weakly decreases. There comes a point when it is no longer worth selecting it.

Proposition 2.3. *For all i in $\{1,2\}$, if $\lambda = 1$, then the unique eqilibrium has each of player 1 and player 2 pick player 1's most preferred leading provision quantity per player.*

This result is intuitive enough. When $\lambda = 1$, each player is committed to do at least as much as her co-player. Player 1 has the lowest costs, and, therefore, wants a higher provision level than player 2. Therefore, unless player 2 contributes a half of that provision level, player 1's most prefered leading quantity, player 1 will take the lead and force her hand.

Lemma 2.2. *For all i in $\{1,2\}$, there exists two thresholds, $\underline{\lambda} \leq \bar{\lambda} < 1$ such that,*

(*i*) *whenever $\lambda \in [\bar{\lambda}, 1)$, no pure strategy equilibrium exists;*

(*ii*) *whenever $\lambda \in [0, \underline{\lambda}]$, a unique pure strategy equilibrium exist and it is outcome equivalent to a Nash equilibrium.*

Lemma 2.2 shows in which sense the equilibrium with $\lambda = 1$ is "saliant." In words, it says that very specific circumstances need to be met for a constant system of mutual demand-rights to bring about an outcome without free riding. Circumstances are met, Proposition 2.3 shows, when the concept of free riding has application to all contributions that lie below the average contribution ($\lambda = 1$). In this event, any player's optimal leading quantity, $\frac{2-\lambda}{\lambda}\bar{x}_{0,j}^{i,L}$, coincides with its minimal acceptable response, $\bar{x}_{0,j}^{i,L}$, as well as with the minimal acceptable response to its minimal acceptable response, $\frac{\lambda}{2-\lambda}\bar{x}_{0,j}^{i,L}$. A potential leader, therefore, is indifferent between the first and the third options and she has no incentives to deviate from her optimal leading quantity when player 2 follows. On the contrary, circumstances arn't met when λ is close to but smaller than 1. In the latter case, going from one's leading contribution, $\frac{2-\lambda}{\lambda}\bar{x}_{0,j}^{i,L}$, to $\frac{\lambda}{2-\lambda}\bar{x}_{0,j}^{i,L}$, the minimal acceptable response to $\bar{x}_{0,j}^{i,L}$, brings about a decrease in the public good quantity and an increase in private consumption that make player i better off, independently of her cost type c_i. Furthermore, if, on the contrary, λ is too small, then even Nash equilibrium contributions are considered legitimate, and the system of demand-rights has nothing to contribute.

The last proposition of this section shows that the existence of the saliant equilibrium

isn't threatened by increases in the number of participants.

Proposition 2.4. *For all i in I, if $\lambda = 1$, then the allocation at which each of player 1 to n pick player 1's most preferred leading provision quantity is a CCE.*

2.4.2 Efficiency and Additive Collective Equilibria

What, concretely, is a mutual system of demand-rights? One possibiilty is conceive of mutual systems of demand-rights as outcomes of what Buchanan named "constitutional changes," that is, changes in the "standards of conduct applicable to all members of the social group" (Buchanan (1962), p. 342). Such a change, he argued, may be expected to occur and last whenever its implementation is unanimously approved of by rational individuals. In other words, when it brings about a Pareto improvement. We have seen the kind of outcomes that constant systems of mutual demand rights are likely to bring about. I now ask under which conditions such an outcome constitutes a Pareto improvement over the Nash outcome.

Proposition 2.5. *For every economy e in $\mathcal{E}^{PG}$, there exists a unique $\bar{\alpha}$ in $(0, 1]$ such that, for every $\alpha \leq \bar{\alpha}$, the α-homogeneization $\tilde{e}$ of e induces a game $G(\tilde{e})$ in which the salient CCE Pareto dominates the Nash Equilibrium.*

In words, while opting for a unitary λ guaranties that the joint intention will be brought about, we may only expect a group of individuals to unanimously approve of a joint commitment to steer clear of free riding when cost heterogeneities are relatively small. Indeed, in the presence of large heterogeneities, high cost types would find themselves worse off bringing their consent to a joint commitment that demands from each to contribute as much as the lowest cost individual. The analysis does exclude the possibility that, *in general*, systems of demand-rights arise in situations with reasonably high levels of heterogeneity. Only, in such situations, these systems will have to be more elaborate. The definition I gave of systems of demand-rights already suggests a possibility. Namely, that demand rights be contingent on individuals' relative contribution costs. Call *a-additive* a system of mutual demand-rights that fulfills the following two conditions:

(i) For all i in $I \setminus \{n\}$, $\lambda(c_i) > \lambda(c_{i+1})$, and

(ii) $\sum_{i \in I} \lambda(c_i) = a$.

The next observation shows that, in the presence of two players, if the system of demand

rights is 2-additive, then non-trivial leading equilibria exist in which high demands are exerted on individual 2.

Observation 2.1. *Consider some economy e in $\mathcal{E}^{PG}$ with $I = \{1, 2\}$ and an a-additive system of demand-rights. Denote $\lambda := \lambda(c_2) \in (0, 1)$ and assume λ is close to but not equal to 1. Demanding $a = 2$ sufficies to guaranty the existence of a leading equilibrium.*

2.5 Conclusion

"In the name of God, Amen. We, whose names are underwritten, the Loyal Subjects of our dread Sovereign Lord King James, by the Grace of God, of Great Britain, France, and Ireland, King, Defender of the Faith, &c. Having undertaken for the Glory of God, and Advancement of the Christian Faith, and the Honour of our King and Country, a Voyage to plant the first Colony in the northern Parts of Virginia; Do by these Presents, solemnly and mutually, in the Presence of God and one another, *covenant and combine ourselves together into a civil Body Politick, for our better Ordering and Preservation, and Furtherance of the Ends aforesaid*: And by Virtue hereof do enact, constitute, and frame, such just and equal Laws, Ordinances, Acts, Constitutions, and Officers, from time to time, as shall be thought most meet and convenient for the general Good of the Colony; unto which we promise all due Submission and Obedience.[22]" Individuals, in their everyday life, have opportunities to jointly agree to act as a body. It is not to be doubted that such opportunities are seized, for instance, when two acquaintances go for a walk, engage in a danse, or converse with one another. In this paper, I defend the view that individuals' genuine attempts to unite their forces are not merely to be observed in familiar and amicable environments, but also in economically relevant ones.

In relatively complex and anonymous situations, such as those involving many players and a public good, a joint commitment to act as a body may be involved too. This is shown by the fact that, in some such instances, participants to the situation use system of demand-rights to check up on each others' behavior. I call *legitimate* those public goods which give rise to a systems of demand-rights. I conjecture that, in their presence, individuals jointly intend to steer clear of free riding. The ambition of this essay, then, was to describe the kind of individual commitments entailed by a joint intention to stear clear of free riding as well as to delineate the type of situations which are likely to give rise to a joint commitment. Situations with low and intermediate levels of heterogeneity,

[22] Mayflower Compact, signed aboard ship on November 21, 1620.

I argue, are good candidates for a joint agreement. The former, despite their propensity to bring about Nash equilibria exempt of free riding, are good candidates because a joint agreement to steer clear of free riding brings about a Pareto improvements when individuals are homogeneous enough. The latter because they tend to yield higher levels of free riding than the former and still may, in the presence of a joint agreement, yield an outcome that Pareto dominates the Nash equilibrium. When heterogeneity is high, equilibrium free riding is high and joint agreements are more complex to implement. Despite higher costs involved, we might expect an institutional solution (Ostrom, 1990) instead.

Chapter 3

Public Good Experiments: a Framing Problem?

"[The general will] must be shown the good road it is in search of, secured from the seductive influences of individual wills [...] The individuals see the good they reject; the public wills the good it does not see."

Jean Jacques Rousseau, *The Social Contract*, 1762.

"[An] alternative is that giving is consistent with social norms about participation in social dilemmas. [...] Decay may simply represent the group's struggles to establish the norm."

James Andreoni, Why Free Ride?, *Journal of Public Economics*, 1988.

3.1 Introduction

An economist willing to follow the official methodological standards of her field, those of revealed preference theory, will make sure to identify individual choices with the satisfaction of individual wants and to remain agnostic about the *contents* of these wants. She will only express statements about wants which are reducible to observation statements[1] and, upon request, may justify her deeds by arguing that, in the realm of science, preference maximization ought to be understood in an "as if" sense. In previous chapters, I argued that the reasons which confine scientists to the "as if" reading of preference

[1] By observation statements I mean statements about observable physical phenomena. For instance, statements to the effect that an individual "prefers x over y" are conventionally reduced to observation statements about physical choices: she chooses x when both x and y are available.

maximization are unsound, and that, when the interactive situation under study involves a legitimate public good, the revealed preference paradigm does not flatten but steepens the hurdle faced by scientific analysis. Indeed, in the presence of a legitimate public good, it is common knowledge that the absence of free riding is necessary to achieve the common good and, for this reason, individuals carry out a collective intention to steer clear of free riding. In this chapter, I consider public good game experiments, which I take to be situations involving a public good whose status falls short of legitimacy. In such situations, too, it is an empirical question whether all individuals follow principles of individual rationality, i.e., maximize their preferences, or whether they abide by a different kind of rationale.

From the physical standpoint, games involve individuals, their available movements, and a set of possible physical consequences. Call *interactive situation* any such collection of individuals, movements, and physical consequences. When involved in an interactive situation, individuals do not apprehend it from the physical standpoint. They ascribe an overarching goal to the situation and they take each involved player to be not merely 'moving' but to be acting *intentionally*, that is, with a view to make some description of the situation come true. In her seminal contribution to the philosophy of action, Anscombe (1957) defends the view that our descriptions of intentional actions neither are expressed in the language of physics nor need be reducible to this language.[2] We consider such events, she suggests, from a teleological standpoint, i.e., we identify involved individuals with goal directed agents and identify their actions with *reasons for acting*. In the presence of linguistic conventions, the set of teleological descriptions which may apply to a given situation is finite. As a consequence, it can be argued that participants *know* something about each other's *possible* intentions and that, against the contentions of revealed preference theorists, an objective distinction *can* be made between interactive situation and their induced *game*—by which I mean, the induced teleological description of the situation by individuals.

Now, the possibility to draw a distinction between an interactive situation and its induced game does not, by itself, give a reason for drawing it. In situations such as one of bilateral trade, where it is common knowledge that all individuals abide by the principles of preference maximization, disregarding the interactive situation and directly setting the analysis at the level of the game appears to be a sensible choice. But it may not be so when we consider public good game experiments; for two reasons. First, public good

[2] "Consider a question 'What is the stove doing?', with the answer 'Burning well' and a question 'What is Smith doing?' with the answer 'Resting.' Would not a *parallel* answer about Smith really be 'breathing steadily' or perhaps 'lying extended on a bed'?" Anscombe (1957), §43, emphasis in the text. See also Anscombe (1958).

experiments constitute, by design, artificial situations. This decreases the likeliness that individuals share a common interpretation of the situation, as they often do in familiar environments. For instance, there is some evidence that, in the case of asymmetric public good games, normative expectations about individual behavior do not make up a social norm (Spiller et al., 2016). Second, while public good experiments give individuals an opportunity to act more or less 'cooperatively'—each individual can unilaterally increase the amount of available public good at the expense of his own stock of private goods—, it is not immediately clear what the meaning of acting 'cooperatively' is, nor whether individuals share a view on that matter or not. In this chapter, I ascribe a very loose meaning to the word cooperating, allow for variations in this meaning across individuals and consider the eventuality that some choose to cooperate while others choose to act individually.

Early theories on the voluntary provision of public goods sought to reduce statements about a potential contributor's wants to statements about (i) the amounts of public and private goods available to him and (ii) the marginal transformation rate characterizing the technology structure. These approaches failed (Ledyard, 1995). A significant fraction of the more recent literature on public good games sticks to the revealed preferences paradigm. That is, only physical quantities are allowed in the domain of agents' preferences and it is still taken for granted that cooperation should be reduced to preference maximizing behavior. Within this part of the literature, fruitful alternatives arise from extentions of the domain of individual preferences: the possibility is considered that individuals may value physical quantities that have to do with others'—as opposed to only their own—well being. Various extensions in the domain of individual preferences have been investigated. Distributional and other-regarding preferences, for instance, exploit the entire allocation of private goods among players, as opposed to a player's individual allocation. Theories of reciprocity take advantage of the fact that players may know each others' action space, they relax the assumption that individual preferences be measurable with respect to realized actions only. Each of these alternatives perform better than traditional, self-regarding preferences. But none of them yields a satisfactory account on its own (Chaudhuri, 2011). Most recent studies along these lines assume heterogeneous populations, i.e., populations of players with differing preference patterns.

These recent developments suggest an alternative possibility. Namely, that, in the context of public good experiments, players do not share a common view regarding the meaning of 'cooperating.' Each individual tags her own and others contributions with one of two labels: *cooperative* or *non-cooperative*, but not all use the exact same rule to do so. In the event that some player tags another player's contribution with the label cooperative,

she views the agent who performed it as a member of the cooperators' team: someone who does her bit. In the other event, she judges the agent who performed the action to be a non-cooperator: someone who reasons strategically. In equilibrium, contributions by each player are performed with a trade-off in mind, that between fostering one's objective, whether collective or individual, and inducing a belief concerning whether she is a team reasoner or a strategic player. The cost of inducing such beliefs is endogenously determined, it coincides with that of the cheapest deviation from one's optimal static play. This ensures that communication, beyond being meaningful, is credible. At the end of every period, after disclosure of information, each player can distinguish between the intention driving optimal static play and deviations away from it. She updates her beliefs about the share of collective reasoners in consequence.

3.2 Related Literature

3.2.1 Public Good Games

Call *parametrization* of a public good game a specification of the number of players, the number of repetitions, the individual endowments in private good, as well as the information availbale at every point in time to each individual. Call *design* of a public good game a specification of the strategy space and of a map from strategy profiles to individual allocations in public and private goods. A reason why choices in public good games are difficult to explain is their high sensibility to variations in the parametrization or design of the game. Standard parametrization involve ten repetitions, four players, identical endowments, and feedback about aggregate contributions at the end of every repetition. Standard designs involve a constant marginal per capita return of the public good below the price of private consumption, yet such that the aggregate return is larger than the per capita return of private consumption.[3] Important stylized facts include: (1) a gradual decrease of average contributions from 40% – 60% in the first period to 20% – 30% in the last one; (2) considerable variation in contributions across individuals and across repetitions. Some contribute all, some nothing, some contribution patterns are monotonic, others not.

Variations in the parametrization and design of public good games have been investigated as well. Several conclusions were drawn; these constitute additional stilized facts.

[3]Here and in the sequel, the aggregate return refers to the marginal per capita return multiplied by the number of players. Thus, the benchmark specification is such that a trade-off exists between individual wills and the general will. Investing in the private good is individually efficient but collectively inefficient, the converse is true of investments in the public good.

Firstly, *average* contributions are increasing in (a) the marginal per capita return of the public good, (b) the number of players, (c) opprotunities for the players to communicate, and (d) opportunities for the players to punish each other. Yet they *need not* increase with (e) individual endowments. Variations in (a) and (b) seem to operate only up to a limit (Laury and Holt, 2008), which suggests that at least a fraction of the population neither abides by other-regarding motives nor seeks to cooperate. Variations in (c) and (d) induce more substantial effects. They can lead to very high contribution rates and even reverse the declining trend described in stylized fact (1) (Fehr and Gächter, 2000). This suggests that information asymmetries might impede coordination and that non-coordinators can be effectively disciplined. Secondly, sorting players into (random) groups has been shown to increase intra-group contributions. The effect is of particular interest for this paper, since it does suggest a link between contributions and players' ability to identify with a group as opposed to identifying with one's individual-self. Thirdly, in the event of an unexpected restart of play at the end of the initially specified number of repetitions, average contributions jump back from low last period levels to higher first period levels. This is known as the *restart effect*.

Usual explanations for these stylized facts identify them to outcomes of a simultaneous exertion of individual wants. But not any kind of wants: wants that are measurable with respect to variations in physical outcomes. In other words, a postulate is made that individuals act *as if* they were concerned exclusively with physical outcomes and employ the means necessary to bring about those outcomes which they judge desirable. When the situation is thought to be a familiar one, the simultaneous exertion of individual wants takes the form of a Nash equilibrium. Else, modellers favor uncoupled learning processes, i.e., learning processes in which each individual's learning rule is independent of the payoff function of other individuals (Hart and Mas-Colell, 2003). A successful model, along these lines, amounts to specifying *simple* individual preferences which, together with the chosen solution concept, predict observed play. Several such specifications have been suggested and tested. *Selfish preferences* postulate that an individual's ranking of outcomes is invariant with respect to all physical variables apart from own private and public consumptions, in which it is increasing. *Distributional preferences* (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) suggest that individuals seek a balanced trade-off between maximized individual material payoff and minimized spread of the overall payoff distribution. *Social welfare preferences* (Charness and Rabin, 2002) attribute altruistic motives to individuals, the utilities of which are monotonically increasing in everyone's consumption levels.

While none of these preference patterns fully explain observed stylized facts, they can

be ranked according to their performance in that dimension (Chaudhuri, 2011). Selfish preferences were tested first and are those which perform the worst. For instance, they fail to explain the persistence of significant contributions in repeated experiments, and, in particular, of those observed in the last period of the game (see, e.g., Ledyard (1995) for an assessment of the failure, and Andreoni and Croson (2008) for mixed evidence on the role of reputation effects in repeated games[4]). Distributional preferences and social welfare preferences perform better. Each provide a simple and natural rationale for the persistence of significant contributions. But the former fails to explain behavior in games where individuals are heterogeneously endowed, because richer players do not appear to contribute more than others (Buckley and Croson, 2006)). And the latter, besides seeming to be informationally very demanding, have been statistically rejected in some experiments (Palfrey and Prisbrey, 1997). As a consequence, most recent approaches drop the assumption that all individuals share a single motive. Arifovic and Ledyard (2012), for instance, provide an explanation of conditional cooperation using a learning model with some selfish individuals and some individuals with distributional concerns.

All approaches mentioned so far share the assumption that players act in accordance with individual and instrumental rationales. Taking note of the difficulties faced by these approaches, some have considered the possibility of integrating *procedural* aspects into preferences. Falk and Fischbacher (2006) and Ambrus and Pathak (2011), for instance, suggest that reciprocity can explain much of the experimental evidence from public good games. Andreoni and Samuelson (2006) look at repeated prisonners' dilemmas and build a model of rational cooperation between two individuals with an arbitrarily small preference for the cooperative action. The work I present here pertains to an alternative tradition, one which seeks the solution in alterations of the individual character of the rationales followed by individuals.[5] Harsanyi (1982), for instance, considers what happens when individuals follow a rule-utilitarian rationale. Sugden (1984) develops a model of moral reciprocity, and Roemer (2010, 2015) one of Kantian optimization. An ambition in writing this esssay is to take into account the fact that, independently of the exact collective rationale that they follow, individuals can be seen as tagging observed contributions with either of the 'cooperative' or 'non-cooperative' labels. Since, eventually, they will do so slightly differently, the model should also take this into account.

[4] At any rate, in finitely repeated games, reputation building effects either requires a population with mixed motives or a population with boundedly rational agents (Kreps et al., 1982).

[5] " To construct an acceptable theory of philantropy one must [...] jettison one of the three [neoclassical] assumptions [...]: publicness, utility maximization and Nash conjectures. [...] The one I believe the most promising is to drop the assumption of utility maximization." (Sugden, 1982). "I propose that we can explain cooperation by observing that players may be *optimizing* in a non-classical (that is, non-Nash) manner." (Roemer, 2015)

3.2.2 Team Reasoning

Game theory is concerned with the analysis of events involving several individuals, various action possibilities for each individual, and physical outcomes ensuing from each action profile realization. I call the collection of these physical entities and of the map from action profiles to physical outcomes an *interactive situation*. An interactive situation is not only described by scientists, but also by the individuals taking part in it. I call *game* a redescription of the interactive situation in terms of the number of involved *agents* (a.k.a., *players*), their respective set of available intentional actions, and the map from action profiles to the agents respective payoffs. Agents (a.k.a., players) are not physical entities but conceptual ones. They are the entities which reason as a single body. Intentional actions, similarly, are not mere physical motions, but means to get a description come true. Finally, and more conventionally, payoffs do not stand for physical utilities neither, they represent individual wants. A game, in this sense, is a *framed* description of some interactive situation. In a frame, several individuals may be described as one or several agent, several action possibilities may be identified with one or several intentional actions, and physical outcomes are judged along a value scheme.

The account of games as framed descriptions raises at least three questions. First, what *kind* of objects are frames? Second, what are the reasons that a rational player could have to *endorse* a specific frame? Third, which *type* of frames are empirically relevant? The first and second questions reach beyond the scope of this literature review, I touched upon them in the first chapter of this work and will here deal with them very briefly. Rational individuals, when *reasoning* about occuring events, do so by placing them into a coherent network of concepts. In the absence of a logically prior *mode of reasoning*, this 'placing' is, arguably, a matter of choosing one network over another. A frame is just one such conceptual network, it is a lense through which the situation is perceived. There is one prominent reason why rational individuals may be willing to endorse one frame but not another. Frames may be judicious in one type of situation but not in another, exactly in the same way that geometrical analysis is convenient for low dimensional problems of analysis but must give place to functional analysis in problems involving a high number of dimensions.[6] As Bacharach puts it, "we might call a mode of reasoning in games *valid* if it is 'success promoting': given any game of some very broad class, it yields only choices which tend to produce success, as measured by game payoffs" (Bacharach (2006), p. 8).

[6] I do not intend to say that such association of descriptions to events is conscious, only that, at some level, it must take place since it is our only way to rationally grasp events. In this respect, I find noteworthy that we often seem to give much thoughts to unfamiliar situations and very little to familiar ones. In the first case, it may not yet be clear which conceptual lense will yield a sensible outcome for dealing with the situation. In the second case, we might have already settled for a lense.

The theory of *team reasoning* is about explicitly formalizing a type of frame which, some argue (see, in particular, Sugden (1993) and Bacharach Bacharach (2006)), is valid for social dilemmas and problems of coordination. Bacharach (2006) suggests the following informal definition: "*Roughly, somebody 'team reasons' if she works out the best feasible combinations of actions for all the members of her team, then does her part in it.*" According to these words, team reasoning is about framing the interactive situation from a specific vantage point: that of a team. It is about giving priority to answers to the question "What should *we* do?" over answers to the question "What should *I* do?" Consider, for instance, the following interactive situation. You and I live under the same roof and must deal with the house-chores. There are 2 individuals and the range of physical actions to be undertaken for the chores to be done can be described, say, by a continuum from 0 (no involvment) to 1 (full involvment). Finally, the range of possible physical outcomes may also be described by a continuum from 0 (untidy) to 1 (tidy). For simplicity, let us assume that, independently of the reasoning mode, individuals frame the action sets as a binary set: $\{\text{Do}, \text{Don't}\}$. A stand must be taken with respect to the way individuals are framed and outcomes assessed. An account following revealed preference could look like this:

	Do	Don't
Do	$2, 2$	$0, 0$
Don't	$0, 0$	$1, 1$

On this account, you and I are two independent players, say, player 1 and 2 respectively. From your standpoint, the event where chores are done is preferable to the event in which they are not if and only if the burden of cleaning isn't carried by only one of the two individuals. Each of the two events involving an unfair allocation of the burden of cleaing belongs to the set of your least prefered events. From my standpoint, a symetric ordinal ranking is observed. When each of us reason individually, two action profiles are rational, (Do, Do) or $(\text{Don't}, \text{Don't})$, but no normative account can say which one of the two ought to be chosen.

If the situation is framed from the team's point of view, two differences occur. First, you and I no longer are independent players, but members of a team, say, member 1 and member 2. Second, events (i.e., action profile realizations) are no longer described by their consequences for us qua individuals, but are described by their consequences for the team. That is, a judgement is passed with regard to the desirability of each respective event *for the team*. Assuming that the team identifies the desirability of an event with its Pareto efficiency, the following table represents a team reasoning frame for this situation:

(Do,Do)	(Do,Don't)	(Don't,Do)	(Don't,Don't)
2	1	0	0

From the point of view of the team, there is a unique maximizer: (Do, Do). This unique maximizer is the rational outcome prescribed by team reasoning. Each of us reason as a member of the team and, in consequence, perform the action which he has to perform for the team to be best off. That is, we share the burden of doing the house-chores.[7]

The above example is ideal in two regards: (i) given aligned interests, Pareto efficiency constitutes a salient objective for the team; and (ii) to compute the best feasible combination of actions for the team, any player must *know* the other player's preferences. The first point remains valid in public good experiments; the second not. In fact, the second point might be a reason why, so far, the literature on team reasoning has, to the best of my knowledge, focused on small population cooperation games with complete information (Prisoner's Dilemma, Hi-Lo game). Here lies a distinction between my work and the existing literature. I analyse the possibility of extending the theory of team reasoning to more anonymous situations. For instance, situations with large populations or assymetric information. In particular, I ask how team reasonners would act in a context of public good provision, and whether team reasoners would act in ways which match the stylized facts registered by the experimental litterature on public goods provision.

3.3 The Model

Public good games, I argued, are framed descriptions of interactive situations. In this section, I formalize this statement. An *interactive situation* is a possible physical situation involving more than one individual. Let $\mathcal{I}$ be the collection of all possible interactive situations, I denote:

$$\mathcal{I} \equiv \{\iota = (I, A, X, g)\}$$

where an interactive situation, ι, is characterized by the number of involved individuals, I, the physical action space, A, the set of physical consequences X, and a mechanism, $g : A \to X$, that maps realized action profiles into consequences.

Rather than physically describing the interactive situation they are involved in, individuals use teleological forms of description for it. That is, they describe it in terms of a *game* and a *solution concept*, which, together, conceptualize the situation from a specific vantage point: the end which they have ascribed to it. The collection of all teleological forms of descriptions constitutes a *language*, which I denote $\mathcal{L}$. Formally:

[7] The curious reader may ask what distinguishes team reasoning from altruism. The difference has been clearly established in the literature (Sugden, 1993). It suffices to note that adding up the respective payoffs while keeping the individual point of view (i.e., Nash reasoning) would be of no help in the coordination problem illustrated by the first table.

$$\mathcal{L} \equiv \Gamma \times \Phi \equiv \{G = (N_G, S_G, \Theta_G, (\pi_n)_{n=1}^{N_G}), (u_n)_{n=1}^{N_G})\} \times \{\phi : \Gamma \to \cup_{G\in\Gamma} 2^{S_G}\ , \forall G, \phi(G) \in S_G\}$$

A game, G, consists of a set of *players*, N_G, a *strategy space*, S_G, a *type space*, Θ_G, a collection of prior beliefs about type profile realizations, $(\pi_n)_{n=1}^{N_G}$, and a collection of payoff functions, $(u_n)_{n=1}^{N_G}$. A solution concept, ϕ, maps games into subsets of their strategy profiles.

Rational individuals may not use language in an arbitrary way. Their may only endorse a form of description that has legal tender in the linguistic communities to which they belong. Call the act of ascribing to an interactive situation ι a teleological description that has legal tender in some linguistic community a *framing* of that situation. Denote Λ the collection of all possible framings, we have

$$\Lambda \equiv \{\lambda : \mathcal{I} \to \mathcal{L}\}.$$

In words, a framing, λ, maps the interactive situation into a linguistic frame. This frame entails both the game and its solution concept. Because the solution concept embodies a specific end, the frame constitutes a teleological description of the situation.

At last, individuals *peform intentional actions*, that is, they (attempt to) carry out physical movements of which they can assert, given the framing they endorse, that they are in accordance with the solution of the game they have framed. I Denote

$$\Sigma \equiv \{\sigma : \cup_{G\in\Gamma} 2^{S_G} \to \Delta(A)\}$$

the set of (possibly stochastic) maps from solution sets to performances, and call any element σ of that set an *implementation policy*. To summarize, when facing an interactive situation, a rational individual frames that situation into a game and a solution concept using a conventional framing she masters. Rationality, finally, commands her to implement the solution dictated by the solution concept. I now parametrize each of these elements so as to fit the kind of situation under analysis: public good experiments.

3.3.1 Interactive Situation

The interactive situation, ι, specifies commonly known physical facts which obtain in the considered situation. These entail the number of individuals, the set of physical actions available to each player, the set of possible physical consequences, and the map from physical actions to physical outcomes. I formalize situations involving a public good as

$$\iota = (I, A, X, g),$$

where I denotes both the number and the set of participating individuals, $A \equiv \times_i A_i \equiv [0, w]^I$, the set of physical actions available to all players, and $X \equiv \mathbb{R}_+^{I+1}$ the physical

outcome space. That is, I assume that, for all i in I, $A_i \equiv [0, w]$, where $w \in \mathbb{R}_+$ and that X is an $I + 1$-dimensional real vector space. The interpretation is that each A_i corresponds to individual i's initial endowment in private good, that all individuals are equally endowed, and that outcomes coincide with the possible allocations of public and private good. Denote $x = (x_0, x_1, ..., x_n)$ a characteristic element of X, then x_0 refers to the amount of public good provided and, for each individual i, x_i is the amount of private good left in his stock. Finally, the mechanism, g, describes the explicit set of rules by which participants to the interactive situation have to abide. Some rules are physical constraints, others are consequences of regulations in place. I focus on linear and voluntary provision mechanisms, that is, mechanisms of the form

$$\begin{aligned} g: \ & A \times \mathbb{R}_+ \to X \\ & ((a_1, ..., a_I), M) \mapsto \left(M \sum_{i=1}^{I} a_i \ , (w - a_i)_{i=1}^{I} \right) \end{aligned}$$

In words, each individual is free to decide which part of his stock to allocate to private consumption and which part to use as a contribution to the public good. The public good production technology is linear, with a marginal rate of transformation equal to M. Much data has been collected in these environments. Stylized facts mentioned above are from such environments.

3.3.2 Game and Solution Concept

When describing the occuring physical events, rational participants in an interactive situation abide by preestablished linguistic conventions. Their descriptions of the situation constitute a frame that they can use as a basis for reasoning. Game theorists usually draw a distinction between two elements of a framed description of the interaction: the *game*, and the *solution concept*. As already mentionned, the game is constituted by a description of the players as well as their respective strategy sets and preferences, that is, value judgements over physical outcomes. The solution concept is a specification of 'the point of the game,' that is, of the set of motives that agents may have upon entering the game. Thereby, it singles out, for every game, a set of strategy profiles which are in line with it.[8]

[8]I do not consider the game to be logically prior to the solution concept. Rather, both are logical pars which must guaranty the existence of a solution, that is, an end, for the game. For a justification of this view, see Chapter 1. One way to gain intuition is by considering card games. In a card game, (associations of) players, possible strategies, and payoffs cannot be specified independently of the overarching goal of the game, and vice versa. For instance, in a 4 participants Bridge, the 4 individuals must form 2 teams of 2, each acting as single players. Conversely, in any cut-throat game each individual is a single player and

I consider a situation in which all players describe the situation using the same stage game and solution concept. In the present case, the considered stage game is a tuple

$$\mathcal{G} = \left(N, S, \Theta, \pi, (u_n)_{n=1}^{N}\right)$$

where $N \equiv \{1, ..., N\}$ denotes both the number and the set of *players*, that is, associations of individuals, $S \equiv \times_{n \in N} S_n$ denotes the *strategy* space, that is, the respective actions which a player could justifiably implement. Θ denotes the type space and $\pi \in \Delta(\Theta)$ a common prior over Θ. Finally, u_n denotes each player's respective comparative evaluation of the outcomes.

It is easiest to start the explicit description of the game with that of Θ, its type space. I interpret the type space as the set of possible associations of players. More precisely, a type ascribed to an individual correspond to one of two roles which the individual may take in the game: that of a *strategic player* or that of a *team member*. All team members *act as a body* and, as a consequence, form a single player. Call this player the *team player*. Each strategic player is a separate additional player. Formally, I assume $\Theta \equiv \times_{i \in I}\{0, 1\} \equiv \{0, 1\}^I$, where an individual has type 0 if he is a team member and 1 if he is a strategic player. The common prior, π, therefore, is a probability distribution over a product space. I assume that types are identically and independently distributed, that is, π is a product probability measure of a single probability distribution over $\{0, 1\}$ which, for convenience, I denote π_0.[9]

Coming to N, the set and number of players. Any $\theta \in \Theta$ can be associated with a diagonal $I \times I$-matrix with binary entries $D_\theta \equiv (d_{i,i} = \theta_i)_{i \in I}$. The product of the vector of individual indexes with this matrix forms an intermediary vector of player indexes, $\tilde{N}_\theta$ in $\{0\} \cup I$. In this vector, team reasoners are ascribed a new 'identity,' index 0, and strategic reasoners preserve their individual identity, index i. Denote $n(\theta, i) \equiv rk(\tilde{N}_i)$ the rank of the ith component of $\tilde{N}_\theta$ in ascending order. Then N, the set of players in the game, is identified with $\{(n(\theta, i))_{i \in I}\}$. Regarding the strategy space, S, I assume that each player can choose between contributing or not to the financing of the public good. Formally, $S \equiv \{0, 1\}^N$. But not every player's contribution has the same impact on the amount of public good provided. Whenever the team is of size I_1, it's contribution is I_1 times more efficient than a contribution by any of the individual players.

This fact is reflected in the payoff structure. Each player n in N disposes of a payoff function u_n. A player's payoff function represents comparative evaluation of all possible

act as such. Not all game-motives combinations are compatible with the existence of a coherent end. For instance, one which specifies teams but requires from each individual that he should never do his bit for the team is self-contradictory.

[9] To carry on the parallel with cards game, types are, in general, distributed according to a degenerate prior, as when David and Rudolf are said to make up one team and Gertrude and Ludwig another.

strategy profiles *from that player's point of view*. A comparative evaluation is a *value judgment*, it orders every strategy profile according to its alignment with certain values. The values which matter are embodied in a payoff function. To say that a team player disposes of a payoff function, therefore, simply amounts to saying that, from a team's point of view, some strategy profiles are judged more or less valuable than others. Denote I_1 the set and number of individuals i such that $\theta_i = 0$ (and, therefore, $n(\theta, i) = 1$). I assume the following:

Assumption 3.1. *For all n in N, $u_n(\cdot)$ maps any (s, θ) in $S \times \Theta$ into*

$$\left[M\left(\sum_{\tilde{n}\in N, \tilde{n}\neq 1} s_{\tilde{n}} + I_1 s_1\right) + 1 - s_n\right] w$$

Note that if each team member were assigned a payoff function similar to that of a strategic player, then the payoff function of the team would be a positive linear transformation of the utilitarian sum of its members' respective utilities.[10] Finally, the framed solution concept is a conventional Bayesian Nash equilibrium, but should be understood as a prescriptive concept rather than a positive one. That is, a solution concept for the stage game is a map from the framed game to its set of rational strategy profiles, rather than a map from the game to the set of physical action profiles.

Definition 3.1. *A prescription of the stage game is a strategy profile*

$$(s_i^*)_{i\in I} \in \{0, 1\}^I$$

such that, for all i in I and all θ in Θ,

$$s_i^* \equiv s^*_{n(\theta,i)}$$

with $s^*_{n(\theta,i)}$ an element of $S_{n(\theta,i)}$ which satisfies, for all $s_{n(\theta,i)}$ in $S_{n(\theta,i)}$,

$$\mathbb{E}_\pi\left[u_{n(\theta,i)}(s^*_{n(\theta,i)}, s^*_{-n(\theta,i)})\right] \geq \mathbb{E}_\pi\left[u_{n(\theta,i)}(s_{n(\theta,i)}, s^*_{-n(\theta,i)})\right]$$

In words, every strategic reasoner i identifies herself to a strategic player $n(\theta, i) \neq 1$. A strategic player takes as given other players' moves, $s_{-n(\theta,i)}$, and selects in response the strategy $s_{n(\theta,i)}$ in $S_{n(\theta,i)}$ which, in expectation, is in line with her values. In this sense, an equilibrium strategy of a strategic player constitutes a prescription which a strategic reasoner *ought to* follow. A team reasoner, instead, identifies with the team player. The team player takes as given other players' moves, s_{-1}, and selects in response strategy s_1 in S_1 which, in expectation, is in line with it's values. The team player's equilibrium

[10] Considering this transformation rather than the sum does not affect orderings of the team, which is what matters for decisions in games. It entails benefits for both calculus and exposition.

strategy constitutes a prescription that each team reasoner *ought to* follow. Expectations are needed because the *realized* size of the team is unknown to all players, team player included. It is important to note the *prescriptive* interpretation: the solution concept does not signle out a subset of physical action profiles, but a subset of intention profiles. If the point of the game is to be fulfilled, individuals ought to carry out physical movements that are in line with the blueprints formed by a selected *intention profile*.[11] Which action counts as fullfiling such intentions is a linguistic, as opposed to game theoretic, question. I turn to it now.

3.3.3 Idiosyncracies in Individual Framings

I assume that, in a public good experiment, the fact that the situation is framed using the above mentioned games and solution concept is common knowledge among rational individuals. The fact that individual descriptions of the situation share a common form, however, does not entail that each event will be interpreted in just the same way by all individuals. In particular, each individual's framing specifies the detailed fashion in which an individual maps physical events into conceptual events and vice versa. Denote Λ the set of all possible framings for the situation at hand, that is, the set of all possible maps between interactive situations referred to as public good experiments and the outlined game–solution-concept pair. Denote λ_i the element of Λ which represents individual i's framing. Individual λ's are identical with respect to all but one dimension. Namely, individuals idiosyncratically tag realized contributions with one of two labels: *cooperative* or *non-cooperative*. Non-cooperative contributions are those that fall below individual i's threshold w_i^λ in $[0, w]$, cooperative contributions are those which lie above individual i's threshold.

Assumption 3.2. *For all i in I, after any realized action profile, the function λ_i maps, any a_j into $\{0, 1\}$ according to the following rule:*

$$a_j \mapsto \begin{cases} 0 & \textit{if } a_j \leq w_i^\lambda, \\ 1 & \textit{else.} \end{cases}$$

In words, individual framings differ in only one aspect: the precise way in which it is assessed whether an action counts as 'contributing' or not. Differences in language types, w_i^λ, captures the eventual variety in individuals' idea of "cooperation." Such

[11] Prior to the framing process, naturally, all individuals are free to act as they please, it is no longer so once every individual has *endorsed* a frame. At this stage, rational agents do have rational obligations derivative of their framing of the situation.

differences may be the reasult of a disagreement with regard to the obligations entailed by a commitment to cooperate. Maybe some individuals like to "take it easy," but others do not. Or they could have to do with the individuals personal characteristics. For instance, individuals may exhibit different degrees of risk aversion and, given the unknown size of the team, perceive the actual contribution cost differently. Combining the two, individuals could be more or less tolerant towards perceptual differences too. The point of the model is to remain agnostic in this respect. Intuitively, differences in framing may matter twice. First, in the performance stage, where individuals implement the solution they have framed. Second, in the updating process that will occur once the game will be transformed into a finitly repeated game.

3.3.4 Performances

In the present framework, game and solution concepts do not yield predictions. Rather, they yield, for each player, a recommended intention with which an individual ought to act if she is to fulfill the obligations entailed by the mode of reasoning she abides by. More concretely, every rational individual performs actions which can be described, according to her own framing, as being in line with her component in the solution outlined for the game. Call implementation policy, $\sigma_i(.)$, a (possibly stochastic) map from the set of solutions of the game, $\phi(G)$, to player i's physical action set, A_i. Denote S^* the solution set and s^* an element of it. I assume

Assumption 3.3. *For all i in I and s^* in S^*, $a_i \overset{\sigma_i}{\sim} [0, w]$ with σ_i such that*

$$supp\left(\sigma_i(s^*)\right) = \begin{cases} [0, w_i^\lambda] & \textit{if } s^*_{n(\theta,i)} = 0, \\ [w_i^\lambda, 1] & \textit{else.} \end{cases}$$

This merely says that *no mistake in performance occurs*. A rational individual with an obligation to fulfill a given intention fulfills that intention. At this stage, no further specification of individual's implementation policies need be given. Eventually, the exact form that these may take will be under discussion when I come to equilibrium selection.

3.3.5 The Repeated Game

This model is about situations where individuals are repeatedly invited to contribute to a public good. In this section, I formalize the equilibrium concept I adopt for such situations. Types are drawn once and for all in period 0 and each individual is privately informed about his own type only. Then, the stage game is played in every of the T

periods $t \in T \equiv \{1, 2, \cdots, T\}$. Before the start of each period $t \in \{2, 3, \cdots, T\}$, an anonymized vector of physical contributions made in the previous period of play is publicly disclosed. With exception of the frame, their own type, and their own history of play, public disclosures are the only informative signals available to the individuals, which I assume to have perfect recall.

In any given period $t \in \{2, 3, \cdots, T\}$, each individual i in I uses the information available to her in order to update her belief π. Note that π is an element of the frame. Therefore, existence of privately observed types and of idiosyncrasies in framing will cause divergences in beliefs. For all $t \in \{2, 3, \cdots, T\}$, denote π_i^t individual i's updated belief. π_i^t depends on the following *individual history*:

$$h_i^t \in H_i^t \equiv N \times S^{t-1} \times (S^{N-1})^{t-1}.$$

In words, an individual history is a collection of facts. A fact about one's individual type, $n(\theta, i)$ in N, is known from period 0 onwards; facts regarding the actions so far carried out by individual i, $(\lambda_i(a_i^\tau))_{1 \leq \tau \leq t-1}$ in S^{t-1}, are updated at the begining of each period; finally, facts regarding anonymized actions of the other players, $\left(\lambda_i(a_j^\tau)\right)_{1 \leq \tau \leq t-1, j \neq i}$ in $(S^{N-1})^{t-1}$, too, are updated at the begining of each period. For convenience, the set of possible initial histories is defined for all i in I as $H_i^1 \equiv N$. For individual i, then, the set of all possible personal histories is

$$\mathcal{H}_i \equiv \bigcup_{t=1}^{T} H_i^t$$

In any given period t in T, every individual i in I has observed a realized history h_i^t. She rationally adjusts the contents of her frame, using a copy of G, denoted G_i^t, which differs from G in at most one aspect: the existence of heterogeneous beliefs over the type space. In fact, I make the following, assumption:

Assumption 3.4. *For all i in I, w_i^λ takes one of two values, $0 < w^{\lambda^L} < w^{\lambda^H} < 1$, with equal chances. Furthermore, for all i in I and all t in T, π_i^t is measurable with respect to h_i^t.*

The measurability assumption is an unawareness assumption: I assume that individuals are unaware of the existence of differences in language. A consequence of this unawareness, in the presence of the assumption of no mistakes in performances, is the possibility to model the updated frame of each individual as a copy of game G with idiosyncratic interim beliefs, $\pi_i^{j,t}$ that are measurable with respect to player types $\theta \in \{0, 1\}^I$. For all i in I, denote $G_i^1 \equiv G$ the first period game and G_i^t, $t \in T \setminus \{1\}$ the subsequent idiosyncratic copies. At every $t > 1$, each individual updates his game framing without being aware that beliefs actually differ along dimensions for which the type is not a sufficient statistic.

The timing of the game looks as follows. There are T periods and each period t contains two stages. In stage one, players interpret the information in accordance with their frame and reframe the situation into a copy G^t of G. They simultaneously infer the relevant prescription for them, $s^{t*}_{n(\theta,i)} \in S_{n(\theta,i)}$. In stage two, players implement that prescription using their implementation policy, σ_i. With regard to prescriptions, I assume the following:

Definition 3.2. *A prescription of the repeated game is a tuple*

$$(s_i^{t*}, \pi_i^t)_{i\in I}^{t\in T}$$

of recommended strategies and interim beliefs such that,

(i) *In every period t, individual i's recommended strategy is a prescription of the stage game G_i^t,*

(ii) *Whenever possible, after disclosure, player i's beliefs are updated using Bayes rule:*

$$\forall i, j \in I, \quad \pi_i^{j,t} \equiv \pi_i^{j,t-1}\,[\cdot \mid h_i^t],$$

3.4 The Voluntary Provision of Public Goods

3.4.1 A Benchmark - Two periods, four individuals, identical framings

It is interesting to first have a look at the predictions of the model in the absence of idyosincracies in language. In this event, assumptions 3 and 5, which guaranty the absence of mistakes in performance, imply that no further discrepencies arise between the individuals' interim beliefs than those entailed by differences in types. Any triple (I, π, M) with I in $\mathbb{N}$, π in $[0, 1]$, and M in $[\frac{1}{I}, 1]$[12] induces a unique repeated game, the prescriptions of which I want to analyse. Most experimental set-ups have individuals sorted in groups of four. I here investigate the simple case in which the number of participants is equal to four and the number of repetitions is two. The following result shows, among other things, that there exists a range of common prior and marginal rates of transformation such that it is rational for strategic reasoners to build a reputation in the first round.

[12] $M \leq 1/I$ is not an interesting case. When M is this low, even a population entirely constituted of completely informed team reasoners would not want to invest in the public good.

Proposition 3.1. *Let $I = 4$ and fix an M in $[1/4, 1]$. The following assertions hold:*

(i) *There is a unique value of the common prior, π_M^l, below which all prescriptions amount to no participation, i.e.,*

$$(s_n^{1*}, s_n^{2*})_{n\in N} = (0,0)_N.$$

(ii) *There are unique values of the common prior, π_M^m and π_M^h, $\pi_M^l \leq \pi_M^m \leq \pi_M^h \leq 1$ such that prescriptions to the effect that only team reasoners should participate in the first period, i.e.,*

$$\left((s_1^{1*}, s_1^{2*}), (s_n^{1*}, s_n^{2*})_{n\in N\setminus\{1\}}\right) \in \{((1,0),(1,0)_{N-1}), ((1,0),(0,0)_{N-1})\},$$

exist if and only if $\pi \in [\pi_M^l, 1] \setminus [\pi_M^m, \pi_M^h]$.

(iii) *There are unique values of the common prior, $\pi_M^{m'}$ and $\pi_M^{h'}$, $\pi_M^l \leq \pi_M^{m'} \leq \pi_M^{h'} \leq 1$ such that prescriptions to the effect that all should participate in the first period, i.e.,*

$$\left((s_1^{1*}, s_1^{2*}), (s_n^{1*}, s_n^{2*})_{n\in N\setminus\{1\}}\right) \in \{((1,1),(1,0)_{N-1}), ((1,0),(0,0)_{N-1})\},$$

exist only if $\pi \in \left[\pi_M^{m'}, \pi_M^{h'}\right]$.

In words, Proposition 1 states that the (π, M) space is entirely covered by four regions. First, a South-West region (dark blue) within which (π, M) pairs are so low that no individual is ever prescribed to contribute. Second, a central region (light blue) within which team reasoners ought to contribute in the first period. This may be for one of two reasons: either because expectations on the team size are high enough, or because information about the team size is worth acquiring. Third, a smaller, and yet not entirely overlapped, central region (green) in which contributions may be prescribed to all individuals. Fourth and last, a region (colorless) in the East–South-East corner of the (π, M) space, where no pure strategy prescriptions exist.

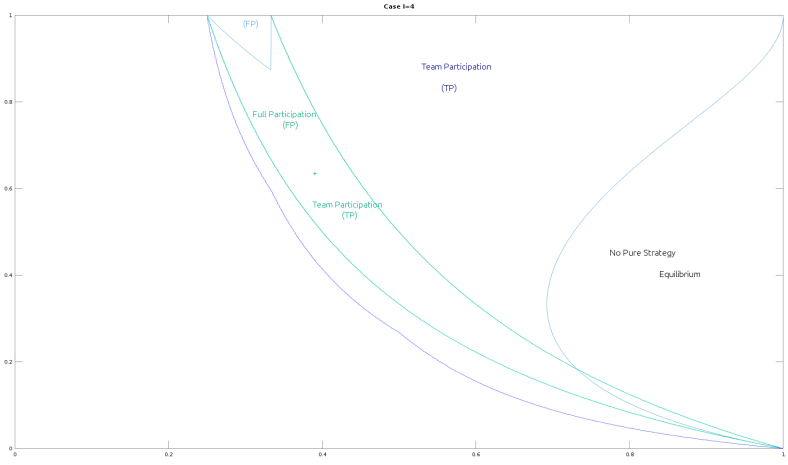


Figure 3.1. Abscissa: M, ordinate: π. π_M^l is in dark blue, π_M^m and π_M^h are in light blue, and $\pi_M^{m'}$ and $\pi_M^{h'}$ are in green.

The absence of a pure strategy prescription in the last region is the consequence of two simultaneous events: the existence of an incentive for strategic players to deviate from non-contributing to contributing when the team player contributes and no strategic player contributes; and the existence of an incentive for strategic players to deviate from a situation in which everyone contributes. The most interesting finding, maybe, is the existence of an interior range of (M, π) values within which reputation building is rational for strategic players. It is a well know result of standard models that, when all players are individually rational, noone has a reason to try to build a reputation in a finitely repeated prisonner's dilemma.

3.4.2 Increase in the Number of Individuals

Formal generalizations of the results to cases with an arbitrary number of players are not straightforward. The reason is that incentive constraints a governed by the probability mass function of a binomial distribution. Numerical results for the cases with 10 and 25 players suggest that (i) the region where only team reasoners are prescribed to participate in the first period grows as I becomes larger, and (ii) all other regions tend to vanish as I increases.

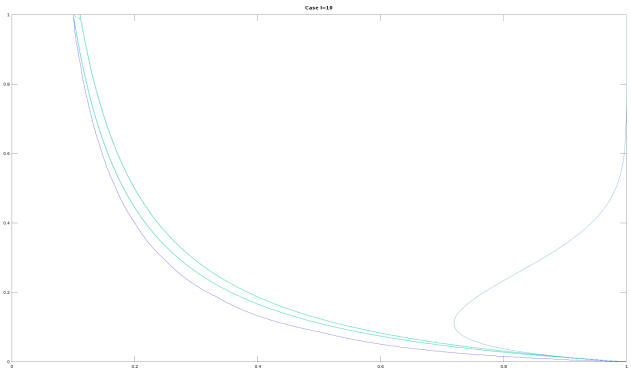


Figure 3.2. Case with 10 players (numerical result).

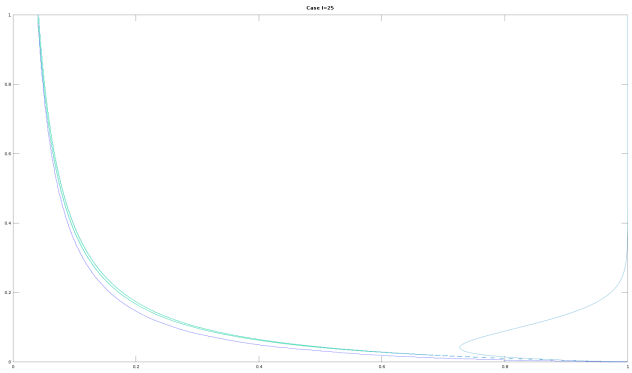


Figure 3.3. Case with 25 players (numerical result).

These observations suggest an explanation why, when the number of participants increases, the number of actually contributing individuals does not significantly go down. On the one hand, in the presence of a large number of players, strategic reasoners loose any influence they could, by deviating, have had on other reasoners' beliefs. They therefore loose all incentives to contribute. On the other hand, given any marginal rate of transformation, the range of beliefs for which it is sensible for team reasoners to contribute becomes significantly larger, potentially large enough to convince pessimistic team reasoners.

3.5 Conclusion

The results so far obtained are broadly in line with experimental findings. They should be subjected to robustness checks along two dimensions: the presence of any finite number of individuals and the possibility that the number of periods be any finite number larger than two. Although numerical simulations I effectuated along the first dimension are encouraging, it is not clear that the results can easily be formally generalized. The reason is that incentive constraints are characterized by inequalities which involve a binomial distribution. It is well known that, when the number of trials is large, such distributions aren't easily computed.

The core of the analysis should focus on idiosyncracies in language. The presence of different thresholds to qualify contributions as a 'cooperative' move or a 'non-cooperative' one will bring about divergences in individual beliefs over time. They may explain, too, why communication can have a significant impact on contributions. I have shown in the second chapter that, in the presence of a joint commitment to stear cleer of free riding, indical individuals can bring about a Pareto efficient situation provided they agree on calling a free rider anyone who contributes less than the average.

Appendix A

Addendum to Chapter 1

A.1 Glossary

Analytic statement: An analytic statement is a statement whose truth or falsity may be assessed by mere study of the (definitional) meaning of the words it contains. "A triangle has three side" is an instance of an analytic statement. Analytic statements are distingusihed from synthetic statements, whose truth or falsity may only be assessed by factual observation.

A posteriori (*knowledge:* knowledge that relates to the results of experience, i.e., the observation of actually realized states of affairs.

A priori (*knowledge*)*:* knowledge that is prior to experience, i.e., invariant to changes in obtaining states of affairs.

Behaviorism: Samuelson and Friedman's methological positions are two instances of behaviorism. The structure of their arguments, as well as Samuelson's formal statements (Samuelson, 1963), prove them to be distinct. Nonetheless, both views have two premisses in common: (1) that a clear, rationally justified, distinction can be operated between observables and unobservables, and (2) that behavior, as opposed to preferences, is observable. They also share an important conclusion: that economists assertions about the world are limited to assertions about choice behavior. This earns them the title of "behaviorism."

Discourse (*forms of*)*:* I adopt here Sellars' view that the logic of discourse is "polydimentional" (see Sellars (1956), esp. §40) and take it to mean that discourse is best seen as a collection of social practices, as opposed to a single one. A form of discourse is one such practice: each utterance of a sentence can be identified with a move—from a context to a sentence, or from one sentence to another—whose adequacy is determined by social

approval or disapproval. Every form of discourse is characterized by a logic, that is, a set of rules characterizing the moves which are approved of. In rare instances the set of rules is explicit; this is the case of *logical discourse*. If I engage in the practice of logic, then each of my moves must distinguish premisses from conclusion and preserve the validity of the argument, i.e., abide by one of the established rules of inference: modus ponens, reductio ad absurum, adjunction, etc. Carroll (1895) pointed out the conventional nature of any agreement to use these rules by pointing out the absence of a compelling *justification* for their use. Wittgenstein (1921) expressed a similar view when identifying logical statements with "showings" as opposed to "sayings." Scientific discourse is an instance of a form of discourse where the rules are, to a large extent, implicit. The rules of inference which prevail in this practice differ from those of logic; they include, for example, causal inference. Hume (1739) (see Book I, part III, section VI) pointed out to a similar problem with the rule of causal inference: one may use it but not justify it.

Discourse (*logic of*)*:* the logic of a form of discource is the set of rules which charaterize that form of discourse. These rules need not be explicit.

Epistemic attitude: epistemology is a field in philosophy concerned with the study of knowledge and justified belief. Attitude is to be understood in it's ordinary sense, i.e., as "a settled way of thinking or feeling about something" (Oxford dictionaries).

Fact: I take facts to be true propositions. It is important to distinguish this use of the word from that which associate it with occuring events.

Mentalism: a philosophical position which holds that coherent accounts of concepts such as 'knowledge,' 'preference,' or 'intention,' cannot dispense with references to inner states of the considered person's mind.

Ontology: the branch of philosophy concerned with what there *is*, i.e., with the set of entities that must be included in our accounts of 'reality.'

Ostensive: an ostensive definition conveys the meaning of a term by pointing out examples. This idea that the meaning of terms is acquired ostensively has a long tradition, as Wittgenstein's quote of Augustin shows: "When grown-named some object and at the same time turned towards it, I perceived this, and I grasped that the thing was signified by the sound they uttered, since they meant to point *it* out. This, however, I gathered from their gestures, the natural language of all peoples, the language that by means of facial expression and the play of eyes, of the movement of the limbs and the tone of voice, indicates the affections of the soul when it desires, or clings to, or rejects, or recoils from, something. In this way, little by little, I learnt to understand what things the words,

which I heard uttered in their respective places in various sentences, signified. And once I got my tongue around these signs, I used them to express my wishes." (Augustine, *Confessions*, I. 8., in Wittgenstein (1953), p. 1)

Observable: observables are the set of perceptible entities in the presence of which and to which observational discourse can be applied. A disagreement subsist regarding whether or not the rules which govern observational discourse are purely physiological or also involve conventions. As far I can see it, a strong movement occured in favor of the latter view in third quarter of the twentieth century.

Practice: I follow (Rawls, 1955) and call practice any "form of activity specified by a system of rules which defines offices, roles, moves, penalties, defenses and so on, and which gives the activity its structure."

Private vs. Public: when it comes to elements of perception, I distinguish private, or inner, elements of perception from public ones. I understand the former as accessible to a specific individual only and the latter as accessible to *any* individual located at the right place in the right moment. An example of the former is a feeling of pain, an example of the latter is an expression of that feeling.

Appendix B

Addendum to Chapter 2

B.1 Proof of Proposition 2.1

For all i in I, define

$$A_i \coloneqq \left\{x_0^{-i} \mid \xi_i(\bar{w} + c_i x_0^{-i}) - x_0^{-i} \geq 0\right\},$$
$$B_i \coloneqq \left\{x_0^{-i} \mid \xi_i(\bar{w} + c_i x_0^{-i}) - x_0^{-i} < 0\right\},$$

The assumption that the public good is a normal good implies that the function conditioned upon is continuous and monotonically decreasing in x_0. Hence, there extists a unique $\bar{x}_0^i$ such that $A_i \equiv [0, \bar{x}_0^i]$ and $B_i \equiv (\bar{x}_0^i, +\infty)$.

For all i in I, define individual and aggregate replacement functions (Cornes and Hartley, 2007) respectively as follows:

$$r_i(x_0, \bar{w}, c_i) \coloneqq \begin{cases} \frac{\bar{w} - \xi_i^{-1}(x_0)}{c_i} + x_0 & \text{if } x_0 \leq \bar{x}_0^i \\ 0 & \text{else} \end{cases}$$

and

$$R(x_0, \bar{w}, c) \coloneqq \textstyle\sum_{i \in I} r_i(x_0, \bar{w}, c_i)$$

The individual replacements functions are continuous and decreasing in x_0 over A_i. They pick up, for any level of public good x_0 produced by players other than i, the unique quantity q such that player i's Nash best response satisfies $BR_i^N(x_0 - q) = q$.

A Nash equilibrium is a strategy profile $\left((x_{0,i}^*, x_i^*)\right)_{i \in I}$ such that, for all i,

$$x_{0,i}^* = r_i(x_0^*, \bar{w}, c_i) \text{ and } x_i^* = \bar{w} - x_{0,i}^*$$

where $x_0^* \equiv \sum_{i \in I} x_{0,i}^*$. The quantity of public good provided at a Nash equilibrium, x_0^*,

coincides with a fixed point of the aggregate replacement function. In equilibrium, the difference between player i and player j's contirbutions is

$$x^*_{o,i} - x^*_{0,j} = r_i(x^*_0, \bar{w}, c_i) - r_j(x^*_0, \bar{w}, c_j)$$

For all i in I, Cobb-Douglas preferences with weight θ on the public good and $1-\theta$ on the private good lead to Engel curves $\xi_i(w) = \frac{\theta w}{c_i}$. Considering now the equilibrium contributions of some individual i, $i < I$, and her nearest more-productive co-player, $i-1$; we have:

$$\begin{aligned} r_{i-1}(x^*_0, \bar{w}, c_{i-1}) - r_i(x^*_0, \bar{w}, c_i) &= \bar{w}\left(\frac{1}{c_{i-1}} - \frac{1}{c_i}\right) + \left(\frac{\xi_i^{-1}(x^*_0)}{c_i} - \frac{\xi_{i-1}^{-1}(x^*_0)}{c_{i-1}}\right) \\ &= \bar{w}\left(\frac{1}{c_{i-1}} - \frac{1}{c_i}\right) \end{aligned}$$

which, clearly, increases with either of (i) an increase in $\bar{w}$, or (ii) a homogeneous decrease in c. To see why (iii) also holds, observe that, the requirements that (a) $\tilde{c}_1 = c_1$ and (b) $\tilde{c}_i - \tilde{c}_{i-1} = \alpha(c_i - c_{i-1})$ together implies that, for all i in I, $\tilde{c}_i = \alpha c_i + (1-\alpha)c_1$. Therefore, when α is sufficiently small,

$$\begin{aligned} \frac{1}{\tilde{c}_{i-1}} - \frac{1}{\tilde{c}_i} &= \frac{\tilde{c}_i - \tilde{c}_{i-1}}{\tilde{c}_{i-1}\tilde{c}_i} \\ &= \frac{c_i - c_{i-1}}{\alpha c_i c_{i-1} + (1-\alpha)c_1\left[c_i + c_{i-1} + \frac{1-\alpha}{\alpha}c_1\right]} < \frac{c_i - c_{i-1}}{c_i c_{i-1}} = \frac{1}{c_{i-1}} - \frac{1}{c_i} \end{aligned}$$

This concludes the proof.

■

B.2 Proof of Proposition 2.2

Any allocation x arrived at through $G(e)$ is characterized by a collection of individual contributions fully ordered by the usual order on $\mathbb{R}$. Consider an individual i who is not the smallest contributor and let j be the individual such that rank$(x_{0,j})$ falls short of rank$(x_{0,i})$ by one unit. Then, $x_{0,j} \leq x_{0,i}$ and if j isn't an eventual free rider, i.e.,

$$x_{0,j} \geq \frac{\lambda}{n - |I^j_-|\lambda} \sum_{l \in I^j_+} x_{0,l}$$

then,

$$
\begin{aligned}
& x_{0,i} \geq \frac{\lambda}{n-(|I_-^i|-1)\lambda}\left(x_{0,i}+\sum_{l\in I_+^i} x_{0,l}\right) \\
\Leftrightarrow \quad & x_{0,i} \geq \left(1-\frac{\lambda}{n-(|I_-^j|-1)\lambda}\right)^{-1} \frac{\lambda}{n-(|I_-^i|-1)\lambda}\sum_{l\in I_+^i} x_{0,l} \\
\Leftrightarrow \quad & x_{0,i} \geq \frac{\lambda}{n-|I_-^j|\lambda}\sum_{l\in I_+^i} x_{0,l} \qquad (\star)
\end{aligned}
$$

In words, if j isn't an eventual free rider, then neither is i. Taking the contrapositive: if equation $(\star)$ does not hold for i, then neither does it for i's predecessor, j. Now, consider an individual i in I whose contribution is such that equation $(\star)$ holds. Multiplying both side of the equation by $\frac{n-|I_-^i|\lambda}{n-(|I_-^i|-1)\lambda} = \left(1-\frac{\lambda}{n-(|I_-^i|-1)\lambda}\right)$ and rearranging, we get

$$
\begin{aligned}
& x_{0,i} \geq \frac{\lambda}{n-(|I_-^i|-1)\lambda}\left(\sum_{j\in I_+^i} x_{0,j}+x_{0,i}\right) \\
\Leftrightarrow \quad & x_{0,i} \geq \frac{\lambda}{n}\left(1+\frac{(|I_-^i|-1)\lambda}{n-(|I_-^i|-1)\lambda}\right)\left(\sum_{j\in I_+^i} x_{0,j}+x_{0,i}\right) \\
\Leftrightarrow \quad & x_{0,i} \geq \frac{\lambda}{n}\left((|I_-^i|-1)\hat{x}+x_{0,i}+\sum_{j\in I_+^i} x_{0,j}\right)
\end{aligned}
$$

with $\hat{x} \equiv \frac{\lambda}{n-(|I_-^i|-1)\lambda}\left(\sum_{j\in I_+^i} x_{0,j}+x_{0,i}\right)$ being the binding contribution of individuals who were brought to meet their obligations prior to i. This shows the equivalence between $(\star)$ and a condition which necessarily and sufficiently characterizes the eventual free riding of individual i.

■

B.3 Proof of Lemma 2.1

I start with an additional Lemma that will be of use in later proofs too.

Lemma B.1. *For any α in $\mathbb{R}_{++}$, denote x_α the value of $x_{0,j}$ that maximizes*

$$u\left((1+\alpha)x_{0,j}, w - c_i\alpha x_{0,j}\right)$$

If $\alpha^1 > \alpha^2$ then $x_{\alpha^1} < x_{\alpha^2}$.

Proof of Lemma B.1

Note that:

(i) $\frac{c_i\alpha}{1+\alpha}$ is increasing in α;

(ii) For each l in $\{1,2\}$, first order conditions must hold at each x_{α^l}:

$$\frac{\partial}{\partial x_0}u\left((1+\alpha^l)x_{\alpha^l}, w - c_i\alpha^l x_{\alpha^l}\right) = \frac{c_i\alpha^l}{1+\alpha^l}\frac{\partial}{\partial x}u\left((1+\alpha^l)x_{\alpha^l}, w - c_i\alpha^l x_{\alpha^l}\right)$$

Assume that $x_{\alpha^1} \geq x_{\alpha^2}$, then

$$\frac{\partial}{\partial x_0}u\left((1+\alpha^1)x_{\alpha^1}, w - c_i\alpha^1 x_{\alpha^1}\right) < \frac{\partial}{\partial x_0}u\left((1+\alpha^2)x_{\alpha^2}, w - c_i\alpha^2 x_{\alpha^2}\right)$$

and

$$\frac{c_i\alpha^1}{1+\alpha^1}\frac{\partial}{\partial x}u\left((1+\alpha^1)x_{\alpha^1}, w - c_i\alpha^1 x_{\alpha^1}\right) > \frac{c_i\alpha^2}{1+\alpha^2}\frac{\partial}{\partial x}u\left((1+\alpha^2)x_{\alpha^2}, w - c_i\alpha^2 x_{\alpha^2}\right)$$

A contradiction.

■

Proof of Lemma 2.1

I take the point of view of some player i in $\{1,2\}$ and proceed by comparing value functions induced by both strategies. That is, I take a look at the $(x_{0,j}, u_i)$-space.

Fix λ and let $\bar{x}^{i,L}_{0,j}(\lambda) \equiv \arg\max_{x_0\in\mathbb{R}_+} u\left(\frac{2}{\lambda}x_0, w - c_i\frac{2-\lambda}{\lambda}x_0\right)$ denote the unconstrained maximum of player i's lead utility function. Given a contribtion $x_{0,j}$ by player j, leading yields

$$u^L(x_{0,j}) = \begin{cases} u\left(\frac{2}{\lambda}\bar{x}^{i,L}_{0,j}(\lambda), w - c_i\frac{2-\lambda}{\lambda}\bar{x}^{i,L}_{0,j}(\lambda)\right) & \text{if } x_{0,j} \leq \bar{x}^{i,L}_{0,j}(\lambda) \\ \\ u\left(\frac{2}{\lambda}x_{0,j}, w - c_i\frac{2-\lambda}{\lambda}x_{0,j}\right) & \text{else,} \end{cases}$$

which is weakly decreasing in $x_{0,j}$.

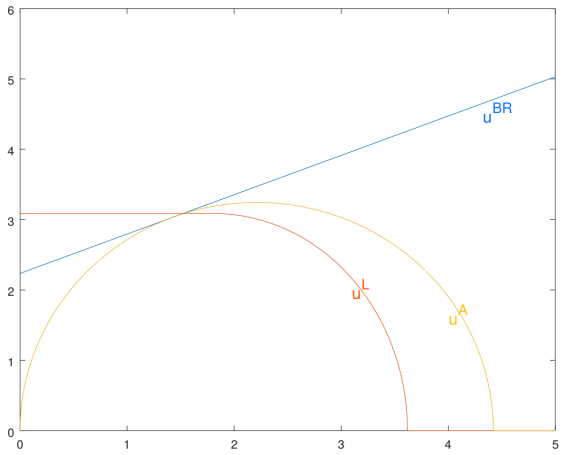
Playing one's adjusted Nash best response, instead, may be shown to yield a payoff that is increasing in $x_{0,j}$ as long as

$$x_{0,j} \leq \bar{x}^{i,A}_{0,j}(\lambda) \equiv \arg\max_{x_0\in\mathbb{R}_+} u\left(\frac{2}{2-\lambda}x_0, w - c_i\frac{\lambda}{2-\lambda}x_0\right)$$

and decreasing afterwards. The reason is twofold:

(a) Call $u^A(x_{0,j}) \equiv u\left(\frac{2}{2-\lambda}x_{0,j}, w - c_i\frac{\lambda}{2-\lambda}x_{0,j}\right)$ player i's payoff from *merely abiding* by the obligations entailed by gift $x_{0,j}$. It follows from concavity and the absence of substitutability between the two goods that $u^A(\cdot)$ is a bell shaped function of $x_{0,j}$. Call $u^{BR}(x_{0,j}) \equiv u\left(BR_i(x_{0,j}) + x_{0,j}, w - c_iBR_i(x_{0,j})\right)$ the utility individual i gains from disregarding the obligation system and simply playing her Nash *best response*. Since, from player i's point of view, an increase in $x_{0,j}$ is tantamount to a virtual increase her wealth, $u^{BR}(\cdot)$ must be increasing in $x_{0,j}$.

(b) Denote $\tilde{x}_{0,j}$ the point that solves $BR_i(\tilde{x}_{0,j}) = \frac{\lambda}{2-\lambda}\tilde{x}_{0,j}$. By defintion, Nash best responding yields a higher payoff to player i than merely abiding, except at point $\tilde{x}_{0,j}$, where the two payoff functions are tangent. Furthermore, for any given λ, the left hand side of the equality that defines $\tilde{x}_{0,j}$ is decreasing in $x_{0,j}$ and the right hand side increasing. So, when facing some $x_{0,j}$ smaller than $\tilde{x}_{0,j}$, player i can choose her best response without breaking her obligations. But once $x_{0,j}$ is larger than $\tilde{x}_{0,j}$, she has an obligation to give up on her Nash best response and provide the minimal acceptable response to $x_{0,j}$, $\frac{\lambda}{2-\lambda}x_{0,j}$. $u^{\tilde{BR}}(\cdot)$, therefore, coincides with $u^{BR}(\cdot)$ when $x_{0,j} \leq \tilde{x}_{0,j}$ and with $u^A(\cdot)$ for larger values of $x_{0,j}$. Tangency at $\tilde{x}_{0,j}$ implies a positive slope at $u^A(\tilde{x}_{0,j})$, so it must be the case that $\tilde{x}_{0,j} < \bar{x}^{i,A}_{0,j}(\lambda)$.



A Cobb-Douglas Case with λ Smaller than Unity.

I first treat the case $\lambda = 1$, which is distinct from the others. Indeed, in this event, $\lambda = 2-\lambda$ and $u^L(\cdot)$ and $u^A(\cdot)$ coincide for every $x_{0,j} \geq \bar{x}^{i,L}_{0,j}(1) \equiv \bar{x}^{i,A}_{0,j}(1)$. Prior to $\bar{x}^{i,L}_{0,j}(1)$, u^L is identically equal to the maximal value of u^A. As a consequence, the graph of the leading value function lies strictly above that of the adjusted best response value function as long as $x_{0,j}$ lies in $[0, \bar{x}^{i,L}_{0,j}(\lambda))$ and the two coincide afterwards. Set $\hat{x}^i_{0,j}(1) \equiv \bar{x}^{i,L}_{0,j}(1)$. It is the case that player i strictly prefers taking the lead over playing her adjusted Nash best response if and only if $x_{0,j} < \hat{x}^i_{0,j}(\lambda)$. Else, she is indifferent.

When $0 < \lambda < 1$, we must make sure that, as in the figure above, $u^L(\cdot)$ and $u^{\tilde{BR}}(\cdot)$ cross exactly once—call this point $\hat{x}^i_{0,j}(\lambda)$—and are such that $u^L(x_{0,j}) > u^{\tilde{BR}}(x_{0,j})$ if and only if $x_{0,j} < \hat{x}^i_{0,j}(\lambda)$. First, observe that, when player i leads, we may equally view her as selecting her most preferred $x_{0,i}$. That is, she solves:

$$\max_{x_0 \geq \frac{2-\lambda}{\lambda} x_{0,j}} u\left(\tfrac{2}{2-\lambda}x_0, w - c_i x_0\right)$$

which, when $x_{0,j}$ is null, amounts to solving

$$\max_{x_0 \geq \mathbb{R}_+} u\left(\tfrac{2}{2-\lambda}x_0, w - c_i x_0\right)$$

As a consequence, for any $\lambda > 0$

$$u^{\tilde{BR}}(0) = \max_{x_0 \in \mathbb{R}_+} u(x_0, w - c_i x_0) < \max_{x_0 \in \mathbb{R}_+} u(\tfrac{2}{2-\lambda}x_0, w - c_i x_0) = u^L(0)$$

By continuity, it must be the case that, for small $x_{0,j}$'s, player i is better off leading.

Second, by Lemma B.1, $\bar{x}^{i,A}_{0,j}(\lambda) \geq \bar{x}^{i,L}_{0,j}(\lambda)$, with a binding inequality if and only if $\lambda = 1$. Thus, $u^L(\cdot)$ and $u^{\tilde{BR}}(\cdot)$ must cross exactly once on the ascending side of $u^{\tilde{BR}}(\cdot)$. Furthermore, for all $x_{0,j} \geq \bar{x}^{i,L}_{0,j}(\lambda)$,

$$\tfrac{d}{dx_{0,j}}\left(u\left(\tfrac{2}{\lambda}x_{0,j}, w - c_i\tfrac{2-\lambda}{\lambda}x_{0,j}\right)\right) < \tfrac{d}{dx_{0,j}}\left(u\left(\tfrac{2}{2-\lambda}x_{0,j}, w - c_i\tfrac{\lambda}{2-\lambda}x_{0,j}\right)\right)$$

Therefore, it is not the case that the two curves cross on the descending side of $u^{\tilde{BR}}(\cdot)$. This completes the proof.

■

B.4 Proof of Proposition 2.3:

Consider an arbitrary player i in $\{1, 2\}$. Player i's utility from playing her adjusted best response is,

$$u^{\tilde{BR}}(x_{0,j}) = \begin{cases} u^{BR}(x_{0,j}) \equiv u\left(BR_i(x_{0,j}) + x_{0,j}, w - c_i BR_i(x_{0,j})\right) & \text{if } x_{0,j} \leq \tilde{x}_{0,j}(\lambda) \\ \\ u^A(x_{0,j}) \equiv u\left(\tfrac{2}{2-\lambda}x_{0,j}, w - c_i\tfrac{\lambda}{2-\lambda}x_{0,j}\right) & \text{else.} \end{cases}$$

and her utility from leading is:

$$\max_{x_0 \geq x_{0,j}} u\left(\tfrac{2}{\lambda}x_0, w - c_i\tfrac{2-\lambda}{\lambda}x_0\right)$$

Each of the two functions is located in the $(x_{0,j}, u_i)$ space; their general shape was described in the proof of Lemma 2.1. I am here interested in their upper envelope.

Assume that $\lambda = 1$. Let $\bar{x}_{0,j}^{i,A}(1)$ be the value of $x_{0,j}$ at which $u^A(\cdot)$ is maximized when $\lambda = 1$. I show first that there is no $x_{0,j}$ at which the two folloing facts simultaneously hold:

(i) individual i's best response is such that she abides by her obligations;

(ii) best responding is making individual i best off.

In other words, whenever the Nash best response is such that individual i fulfills her obligations, individual i is better off taking the lead. To see why, observe that, when $\lambda = 1$, the objective function of the constrained maximization problem that player i has to solve when she leads is identically equal to $u^A(\cdot)$. Consequently, the upper envelope of $u^L(.)$ and $u^A(.)$ coinides with $u^L(x_{0,j}) = u^A(\bar{x}_{0,j}^{i,A}(1))$ between the origin and $\bar{x}_{0,j}^{i,A}(1)$ and is identically equal to $u^A(x_{0,j})$ afterwards. Now, from the proof of Lemma 2.1, we know that $u^{BR}(\cdot)$ is increasing in $x_{0,j}$, always above $u^A(.)$, and tangent to $u^A(.)$ at $\tilde{x}_{0,j}$, the point that solves $BR_i(\tilde{x}_{0,j}) = \tilde{x}_{0,j}$. Furthermore, $\tilde{x}_{0,j}$ must be strictly smaller than $\bar{x}_{0,j}^{i,A}(1)$, because tangency is characterized by identity of slopes. Therefore, $u^{BR}(x_{0,j})$ is smaller than $u^L(x_{0,j})$ whenever $x_{0,j} \leq \tilde{x}_{0,j} < \bar{x}_{0,j}^{i,A}(1)$, and such that player i does not abide by her obligations afterwards.

It follows that each player i in $\{1, 2\}$ opts for her leading strategy as long as $x_{0,j} \leq \bar{x}_{0,j}^{i,A}(1)$ and is indifferent between leading and following afterwards. $x_{0,j}^{i,A}(1)$ solves:

$$\frac{\partial}{\partial x_0}\left(u(2x_{0,j}^{i,A}(1), w - c_i x_{0,j}^{i,A}(1))\right) = \frac{c_i}{2}\frac{\partial}{\partial x}\left(u(2x_{0,j}^{i,A}(1), w - c_i x_{0,j}^{i,A}(1))\right)$$

and is therefore decreasing in c_i, player i's cost for providing the public good. From $c_2 > c_1$ it follows that, when player 1 picks $\bar{x}_{0,2}^{1,A}(1)$, player 2 will accept to follow, i.e., to contribute $x_{0,2} = \bar{x}_{0,2}^{1,A}(1)$. Finally, when player 2 makes such a contribution, player 1 cannot profitably deviate from $\bar{x}_{0,2}^{1,A}(1)$. The converse is not true. This completes the proof.

■

B.5 Proof of Lemma 2.2

Proof of Part (i)

Assume now that $0 < \lambda < 1$. It is clear from the proof of Lemma 2.1 that $u^L(\cdot)$ intersects each of $u^{BR}(\cdot)$ and $u^A(\cdot)$ exactly once. Given any λ, denote $x_{0,j}^{i,BR}(\lambda)$ the intersection between $u^L(\cdot)$ and $u^{BR}(\cdot)$ and $x_{0,j}^{i,A}(\lambda)$ the intersection between $u^L(\cdot)$ and $u^A(\cdot)$. By defini-

tion, $u^{BR}(\cdot)$ lies above $u^{A}(\cdot)$ everywhere but at $\tilde{x}_{0,j}(\lambda)$. Since $u^{L}(\cdot)$ is a decreasing fucntion of $x_{0,j}$, it must be that, for any λ, $x_{0,j}^{i,A}(\lambda) \geq x_{0,j}^{i,BR}(\lambda)$.

Our point of interest is $\hat{x}_{0,j}^{i}$, the intersection of $u^{L}(\cdot)$ and $u^{\tilde{BR}}(\cdot)$. I first show that there exists $\bar{\lambda}_i$ in $(0,1)$ such that,

$$\text{if } \lambda \geq \bar{\lambda}_i, \text{ then } \hat{x}_{0,j}^{i} = x_{0,j}^{i,A}(\lambda) > x_{0,j}^{i,BR}(\lambda)$$

That is, when $\lambda \geq \bar{\lambda}_i$, it is once more the case that

(i) individual i's best response is such that she abides by her obligations, and

(ii) best responding is making individual i best off,

are two mutually excluding states of affairs. To see why, consider the following two thresholds:

$$\tilde{x}_{0,j}(\lambda), \text{ such that } BR_i(\tilde{x}_{0,j}) = \tfrac{\lambda}{2-\lambda}\tilde{x}_{0,j},$$

and

$$\check{x}_{0,j}(\lambda), \text{ such that } BR_i(\check{x}_{0,j}) = \tfrac{2-\lambda}{\lambda}\check{x}_{0,j}.$$

$\tilde{x}_{0,j}(\lambda)$ is the level of $x_{0,j}$ at which $u^{A}(\cdot)$ and $u^{BR}(\cdot)$ are tangent, $\check{x}_{0,j}(\lambda)$ that at which the objective function associated with i's leading strategy and $u^{BR}(\cdot)$ are tangent. *Observe that $\hat{x}_{0,j}^{i}$ coincides with $x_{0,j}^{i,A}(\lambda)$ if and only if $\hat{x}_{0,j}^{i}$ is larger than $\tilde{x}_{0,j}(\lambda)$.* I show that, for λ high enough, $\hat{x}_{0,j}^{i}$ is larger than $\tilde{x}_{0,j}(\lambda)$.

Since (i) $BR_i(\cdot)$ is unaffected by λ and decreasing in $x_{0,j}$, (ii) $\frac{\lambda}{2-\lambda}$ is increasing in λ, and (iii) $\frac{2-\lambda}{\lambda}$ is decreasing in λ, it must be that $\tilde{x}_{0,j}(\lambda)$ and $\check{x}_{0,j}(\lambda)$ are respectively decreasing and increasing in λ. Furthermore, the two coincide when $\lambda = 1$, so we have that, for any λ in $(0,1)$, $\check{x}_{0,j}(\lambda) < \check{x}_{0,j}(1) = \tilde{x}_{0,j}(1) < \tilde{x}_{0,j}(\lambda)$.

We know that $u^{BR}(\cdot)$ is an increasing function of $x_{0,j}$. In addition, we know that $u\left(\frac{2}{\lambda}x_{0,j}, w - c_i\frac{2-\lambda}{\lambda}x_{0,j}\right)$ and $u\left(\frac{2}{2-\lambda}x_{0,j}, w - c_i\frac{\lambda}{2-\lambda}x_{0,j}\right) =: u^{A}(\cdot)$ are bell shaped and identically equal when $\lambda = 1$. Observe that

$$\frac{\partial u}{\partial \lambda}\left(\frac{2}{\lambda}x_{0,j}, w - c_i\frac{2-\lambda}{\lambda}x_{0,j}\right) = \frac{2x_{0,j}}{\lambda^2}\left[c_i\frac{\partial u}{\partial x}\left(\frac{2}{\lambda}x_{0,j}, w - c_i\frac{2-\lambda}{\lambda}x_{0,j}\right) - \frac{\partial u}{\partial x_0}\left(\frac{2}{\lambda}x_{0,j}, w - c_i\frac{2-\lambda}{\lambda}x_{0,j}\right)\right]$$

is negative for low values of $x_{0,j}$ increasing in $x_{0,j}$, and positive after some threshold $x_{0,j}$ lower than the maximizer. On the other hand,

$$\frac{\partial u}{\partial \lambda}\left(\frac{2}{2-\lambda}x_{0,j}, w - c_i\frac{\lambda}{2-\lambda}x_{0,j}\right) =$$

$$\frac{2x_{0,j}}{(2-\lambda)^2}\left[\frac{\partial u}{\partial x_0}\left(\frac{2}{2-\lambda}x_{0,j}, w - c_i\frac{\lambda}{2-\lambda}x_{0,j}\right) - c_i\frac{\partial u}{\partial x}\left(\frac{2}{2-\lambda}x_{0,j}, w - c_i\frac{\lambda}{2-\lambda}x_{0,j}\right)\right]$$

is positive for low values of $x_{0,j}$, decreasing in $x_{0,j}$, and negative after some threshold $x_{0,j}$ lower than the maximizer. Furthermore, Lemma B.1 shows that, given any λ in $(0,1)$,

$$\bar{x}^{i,L}_{0,j}(\lambda) \coloneqq \arg\max_{x_0} u\left(\frac{2}{\lambda}x_0, w - c_i\frac{2-\lambda}{\lambda}x_0\right)$$
$$< \arg\max_{x_0\in\mathbb{R}_+} u\left(\frac{2}{2-\lambda}x_0, w - c_i\frac{\lambda}{2-\lambda}x_0\right) =: \bar{x}^{i,A}_{0,j}(\lambda)$$

Therefore, when $\lambda < 1$, $u\left(\frac{2}{\lambda}\cdot, w - c_i\frac{2-\lambda}{\lambda}\cdot\right)$ is larger than $u^A(\cdot)$ whenever $x_{0,j}$ is small enough, it intersects $u^A(\cdot)$ on its ascending side, and it remains below it for larger $x_{0,j}$s.

Observe that the point at which $u^A(\cdot)$ becomes larger than $u\left(\frac{2}{\lambda}\cdot, w - c_i\frac{2-\lambda}{\lambda}\cdot\right)$ must lie between $\check{x}_{0,j}(\lambda)$ and $\tilde{x}_{0,j}(\lambda)$. Furthermore, when λ is sufficiently close to unity, the following two facts hold:

(a) $\bar{x}^{i,L}_{0,j}(\lambda)$, which lies in a small neighborhood of $\bar{x}^{i,L}_{0,j}(1)$, is larger than $\tilde{x}_{0,j}(1)$. Additionally, each of $\check{x}_{0,j}(\lambda)$ and $\tilde{x}_{0,j}(\lambda)$, which lie in a small neighborhood of $\tilde{x}_{0,j}(1)$, must also be smaller than $\bar{x}^{i,L}_{0,j}(\lambda)$.

(b) $u(\frac{2}{\lambda}\bar{x}^{i,L}_{0,j}(\lambda), w - c_i\frac{2-\lambda}{\lambda}\bar{x}^{i,L}_{0,j}(\lambda))$ is larger than $u^A\left(\tilde{x}_{0,j}(\lambda)\right)$, which lies in a small neighborhood of $u^A(\tilde{x}_{0,j}(1)) < u^A(\bar{x}^{i,L}_{0,j}(1))$.

It follows that

$$u^L(x_{0,j}) = \begin{cases} u\left(\frac{2}{\lambda}\bar{x}^{i,L}_{0,j}(\lambda), w - c_i\frac{2-\lambda}{\lambda}\bar{x}^{i,L}_{0,j}(\lambda)\right) & \text{if } x_{0,j} \le \bar{x}^{i,L}_{0,j}(\lambda) \\ \\ u\left(\frac{2}{\lambda}x_{0,j}, w - c_i\frac{2-\lambda}{\lambda}x_{0,j}\right) & \text{else,} \end{cases}$$

intersects $u^{\tilde{BR}}(\cdot)$ after $\tilde{x}_{0,j}(\lambda)$ has been reached, as claimed.

Now, let $\bar{\lambda} \equiv \max\{\bar{\lambda_1}, \bar{\lambda_2}\}$. Whenever λ lies in $(\bar{\lambda}, 1)$, player i prefers to lead as long as j contributes an amount lower than $x^{i,A}_{0,j}(\lambda)$ and opts for following afterwards. Assume there is a first stage equilibrium in which, say, player j follows player i. That is, player j contributes $\bar{x}^{i,L}_{0,j}(\lambda)$ and player i contributes $\frac{2-\lambda}{\lambda}\bar{x}^{i,L}_{0,j}(\lambda)$. Player i, then, gets a payoff equal to

$$u\left(\frac{2}{\lambda}\bar{x}^{i,L}_{0,j}(\lambda), w - c_1\frac{2-\lambda}{\lambda}\bar{x}^{i,L}_{0,j}(\lambda)\right)$$

which, we know, is smaller than $u^A(\bar{x}^{i,L}_{0,j}(\lambda))$ because the interstection between $u\left(\frac{2}{\lambda}x_0, w - c_i\frac{2-\lambda}{\lambda}x_0\right)$ is reached prior to $\bar{x}^{i,L}_{0,j}(\lambda)$. But $u^A(\bar{x}^{i,L}_{0,j}(\lambda))$ is precisely the utility player i would get from minimally responding to $\bar{x}^{i,L}_{0,j}(\lambda)$. This contradict the assumption that

leading is an equilibrium strategy and concludes the proof of Lemma 2.1 (i).

Proof of Part (ii)

For λ small enough, the following inequalities hold $\bar{x}^{i,L}_{0,j}(\lambda) < \check{x}_{0,j}(1) = \tilde{x}_{0,j}(1) < \tilde{x}_{0,j}(\lambda)$. Thus, $\hat{x}^i_{0,j}$, the intersection of $u^L(\cdot)$ and $u^{\tilde{BR}}(\cdot)$ must coincide with $x^{i,BR}_{0,j}(\lambda)$, the intersection of $u^L(\cdot)$ and $u^{BR}(\cdot)$. Furthermore, by defintion,

$$u\left(\bar{x}^{i,L}_{0,j}(\lambda) + \tfrac{2-\lambda}{\lambda}\bar{x}^{i,L}_{0,j}(\lambda), w - c_i\tfrac{2-\lambda}{\lambda}\bar{x}^{i,L}_{0,j}(\lambda)\right) \leq u\left(\bar{x}^{i,L}_{0,j}(\lambda) + BR_i(\bar{x}^{i,L}_{0,j}(\lambda)), w - c_i BR_i(\bar{x}^{i,L}_{0,j}(\lambda))\right),$$

So it also has to be the case that the two intersect prior to $\bar{x}^{i,L}_{0,j}(\lambda)$, that is, on the constant part of $u^L(\cdot)$. As a consequence, $x^{i,BR}_{0,j}(\lambda)$ solves:

$$u\left(x^{i,BR}_{0,j}(\lambda) + BR_i(x^{i,BR}_{0,j}(\lambda)), w - c_i BR_i(x^{i,BR}_{0,j}(\lambda))\right) = u\left(\frac{2}{\lambda}\bar{x}^{i,L}_{0,j}(\lambda), w - c_i\frac{2-\lambda}{\lambda}\bar{x}^{i,L}_{0,j}(\lambda)\right) \quad \text{(B.1)}$$

The left hand side, when considered a function of $x_{0,j}$, is increasing in its argument. The right hand side, when considered a function of λ, is increasing in its argument too (abstracting from equilibrium requirements, a higher constraining power may only make a leader better off). In consequence, the solution to (B.1) is increasing in λ. Note also that, in the limit case where $\lambda = 0$, the leading utility of player i is $u(\bar{x}^{i,L}_{0,j}(0), w - c_i\bar{x}^{i,L}_{0,j}(0))$, which, by defintion, is identical to $u\left(BR_i(0), w - c_i BR_i(0)\right)$.

Now, let $(x^*_{0,1}, x^*_{0,2})$ denote the unique Nash equilibrium provision levels of the game. In an interior equilibrium, each must be respectively larger than $0 = x^{1,BR}_{0,2}(0) = x^{2,BR}_{0,1}(0)$. As long as λ is such that

$$\mathbb{1}_{\left\{\max\left\{x^{1,BR}_{0,2}(\lambda), x^{2,BR}_{0,1}(\lambda)\right\} \leq \min\left\{x^*_{0,1}, x^*_{0,2}\right\}\right\}}\mathbb{1}_{\left\{|x^*_{0,1} - x^*_{0,2}| \leq \left(1-\frac{\lambda}{2-\lambda}\right)\max\left\{x^*_{0,1}, x^*_{0,2}\right\}\right\}} = 1,$$

each of the two players switches to her best response before Nash equilibrium quantities are reached and the Nash equilibrium, which does not involve breaks of commitments, obtains.

■

B.6 Proof of Proposition 2.4

I proceed directly, i.e., by showing that, at our candidate equilibrium $(x^*, 1)$, each of conditions (i) and (ii) in the definition of Constant Collective Equilibria is fulfilled.

To start with condition (i); at our candidate equilibrium, all players contribute the same amount, namely

$$\bar{x}_{0,i}^{1,L} \coloneqq \arg\max_{x_{0,i}\in\mathbb{R}_+} u\left(nx_{0,i}, w - c_1 x_{0,i}\right)$$

Therefore, for all i in I, $I_-^i = \{i\}$ and $I_+^i = I \setminus \{i\}$. Therefore, equation $(\star)$ is trivially satisfied.

Coming to condition (ii); Recall that, when λ is unitary, the objective function of the maximization problem associated with player j's leading strategy coincides with the function that describes, for every $x_{0,-j}$, the payoff for j from merely abiding by her obligations. At our candidate equilibrium, this payoff is:

$$u_j^A(x_{0,-j}^*) = u\left(n\bar{x}_{0,i}^{1,L}, w - c_j\bar{x}_{0,i}^{1,L}\right)$$

Clearly, individual j cannot be made better of by deviating downwards from $\bar{x}_{0,i}^{1,L}$ because the legitimate adjustment of $\tilde{x} \equiv (\bar{x}_{0,i}^{1,L}, \cdots, \bar{x}_{0,i}^{1,L}, \tilde{x}_{0,j}, \bar{x}_{0,i}^{1,L}, \cdots, \bar{x}_{0,i}^{1,L})$ is equal to x^* whenever $\tilde{x}_{0,j} \leq \bar{x}_{0,i}^{1,L}$. Could her situation be improved by an increase in $x_{0,j}$? It will be the case if and only if

$$BR_j\left((n-1)\bar{x}_{0,i}^{1,L}\right) > \bar{x}_{0,i}^{1,L},$$

or,

$$\bar{x}_{0,i}^{j,L} \coloneqq \arg\max_{x_{0,i}\in\mathbb{R}_+} u\left(nx_{0,i}, w - c_j x_{0,i}\right) > \bar{x}_{0,i}^{1,L}$$

But player 1 is the one with the lowest cost, so neither of these two conditions can ever be fulfilled for some j in I.

■

B.7 Proof of Proposition 2.5

Fix an economy e in $\mathcal{E}^{PG}$ and consider the associated game $G(e)$. Observe that, in a CCE, all individuals contribute

$$x_{0,1}^* \equiv \arg\max_{x\in\mathbb{R}_+} u(nx, \bar{w} - c_1 x)$$

Individual i, therefore, gets utility $u(nx_{0,1}^*, \bar{w} - c_i x_{0,1}^*)$, which, clearly, is decreasing in c_i. We want to compare it to her Nash equilibrium utility. In an interior Nash equilibrium, when individuals have identical preferences, it does not pay to have a comparative advantage at producing the public good: higher cost individuals enjoy a higher utility level (Cornes and Hartley (2007), proposition 4.2). Denote $x^N = (x_{0,1}^N, x_{0,2}^N, \cdots, x_{0,n}^N)$ the Nash equilibrium contributions. Therefore, it must be the case that, for all i in I,

$$u(nx^*_{0,1}, \bar{w} - c_i x^*_{0,1}) - u\left(\sum_{j\in I} x^N_{0,j}, w - c_i x^N_{0,i}\right) \leq u(nx^*_{0,1}, \bar{w} - c_1 x^*_{0,1}) - u\left(\sum_{j\in I} x^N_{0,j}, w - c_1 x^N_{0,1}\right)$$

and

$$u(nx^*_{0,1}, \bar{w} - c_i x^*_{0,1}) - u\left(\sum_{j\in I} x^N_{0,j}, w - c_i x^N_{0,i}\right) \geq u(nx^*_{0,1}, \bar{w} - c_n x^*_{0,1}) - u\left(\sum_{j\in I} x^N_{0,j}, w - c_n x^N_{0,n}\right)$$

In words, individual 1 is the one who gains the most from the constitutional change and individual n the one who gains the least. Unanimity will fail to occur if and only if individual n does gain something, i.e.,

$$u(nx^*_{0,1}, \bar{w} - c_n x^*_{0,1}) < u\left(\sum_{j\in I} x^N_{0,j}, w - c_n x^N_{0,n}\right)$$

Note that, for $c < \tilde{c}$, $\arg\max_{x\in\mathbb{R}_+} u(nx, \bar{w} - cx) > \arg\max_{x\in\mathbb{R}_+} u(nx, \bar{w} - \tilde{c}x)$. Therefore, an α-homogeneization of c will bring about an increase on the left hand side. On the right hand side, the effect is ambiguous. On the one hand, the α-homogeneization is beneficial to individual n because it brings about an increase in the Nash equilibrium provision. On the other hand, it is detrimental to her because she has to increase her contribution to the public good. At any rate, when α goes to 0, the left hand side goes to $u(nx^*_{0,1}, \bar{w} - c_1 x^*_{0,1})$ and the right hand side to $u(nx^N_{0,1}, \bar{w} - c_1 x^N_{0,1})$. The latter is smaller than the former, a continuity argument suffices to conclude the proof.

■

B.8 Proof of Observation 2.1

If $\lambda(c_1) = 2 - \lambda(c_2)$, then player 1 has an obligation to contribute

$$x_{0,1} \geq \frac{\lambda(c_1)}{2}(x_{0,1} + x_{0,2}) = \frac{2-\lambda(c_2)}{\lambda(c_2)} x_{0,2}$$

In words, letting $\lambda \equiv \lambda(c_2) \in (0, 1)$, we find ourselve in a situation in which: (i) Player 2 non longer has an option to lead, but may *enforce* leadership by player 1 whenever the latter contributes a low amount; and (ii) Player 1 still has an option to lead, but may no longer follow. At $\bar{x}^{1,L}_{0,2}(\lambda)$, it is already the case that a best response by Player 1 would have her break her obligations. Since Player 1's utility from best-responding is tangent to her leading utility on the ascending side, she never finds it interesting to best-respond. Now, looking again at $\bar{x}^{1,L}_{0,2}(\lambda)$, player 2 would enjoy a lower provision level but cannot bring it about because she must at least follow. Since her provision cost is greater than that of individual 1, she does not want to deviate upwards either. This concludes the proof.

■

Appendix C

Addendum to Chapter 3

C.1 Proof of Proposition 3.1

Lemma C.1. *Let $I = 4$ and fix $M \in [\frac{1}{4}, 1)$. Let f and g be two functions from $[0,1]$ to $\mathbb{R}$ assigning to any π the respective values:*

$$f_{I,M}(\pi) = 2 - \Big(2 + (I-1)\pi + \tilde{\mathbb{E}}_\pi[\mathbb{1}_{\{I_1 M \geq 1-M\}} I_1]\Big) M,$$

and

$$g_{I,M}(\pi) = (1-M)\tilde{\mathbb{P}}_\pi(I_1 M < 1 - M).$$

Their graphs have exactly one element in common.

Proof of Lemma C.1:

By definition,

$$\tilde{\mathbb{E}}_\pi[\mathbb{1}_{\{I_1 M \geq 1-M\}} I_1] \coloneqq \sum_{l=\lceil (1-M)/M \rceil}^{I-1} \binom{I-1}{l} l \pi^l (1-\pi)^{I-1-l}$$

$$\tilde{\mathbb{P}}_\pi(I_1 M < 1 - M) \coloneqq \sum_{l=0}^{\lceil (1-M)/M \rceil - 1} \binom{I-1}{l} \pi^l (1-\pi)^{I-1-l}$$

Note first that, for any admissible I and M, $f_{I,M}(\pi) = 2(1-M)$, $g_{I,M}(0) = 1-M$, $f_{I,M}(1) < 0$, and $g_{I,M}(\pi) = 0$. Therefore, it suffices to show that the derivative of their difference or the difference of their derivatives has a constant sign. This can be done for each of three ranges whitin which M may lie when we are in the presence of four individuals.

Case 1: $M \in [\frac{1}{2}, 1)$, that is, $\lceil (1-M)/M \rceil = 1$

In this event, $\tilde{\mathbb{E}}_\pi[\mathbb{1}_{\{I_1 M \geq 1-M\}} I_1] = 3\pi$ and $\tilde{\mathbb{P}}_\pi(I_1 M < 1 - M) = (1-\pi)^3$. Thus,

$$\frac{\partial f(\pi)}{\partial \pi} = -6M < -3(1-M)(1-\pi)^2 = \frac{\partial g(\pi)}{\partial \pi}$$

where the inequality follows from the fact that $M \geq \frac{1}{2}$ and $\pi \in [0,1]$.

Case 2: $M \in [\frac{1}{3}, \frac{1}{2})$, that is, $\lceil (1-M)/M \rceil = 2$

In this event, $\tilde{\mathbb{E}}_\pi[\mathbb{1}_{\{I_1 M \geq 1-M\}} I_1] = 6\pi^2 - 3\pi^3$ and $\tilde{\mathbb{P}}_\pi(I_1 M < 1-M) = 1 + 2\pi^3 - 3\pi^2$. Thus,

$$f(\pi) - g(\pi) = 1 - M - 3\pi M - 3(2+3M)\pi^2 - (2+M)\pi^3$$

and

$$\tfrac{\partial}{\partial \pi}\Big(f(\pi) - g(\pi)\Big) = -3M(\pi^2 + 6\pi + 1) - 6\pi(\pi+2) < 0$$

Case 3: $M \in [\frac{1}{4}, \frac{1}{3})$, that is, $\lceil (1-M)/M \rceil = 3$

In this event, $\tilde{\mathbb{E}}_\pi[\mathbb{1}_{\{I_1 M \geq 1-M\}} I_1] = 3\pi^3$ and $\tilde{\mathbb{P}}_\pi(I_1 M < 1-M) = 1 - \pi^3$. Thus,

$$\tfrac{\partial f(\pi)}{\partial \pi} = -3M - 9M\pi^2 = -3M(1+3\pi^2) < -(1-M)(1-3\pi^2) = \tfrac{\partial g(\pi)}{\partial \pi}$$

where the inequality follows from the fact that $M \in [\frac{1}{3}, \frac{1}{2})$ and $\pi \in [0,1]$.

■

Proof of Proposition 3.1:

I proceed by backward induction and, therefore, start with each player $n(\theta, i)$'s period 2 prescription given any updated belief $\pi^2_{n(\theta,i)} \equiv \pi_i[\cdot \mid \theta_i, a^1]$. These take one of two forms:

$$\text{For all } i \text{ in } I \text{ such that } \theta_i = 0,\ s^{2*}_{n(\theta,i)} = \begin{cases} 1 & \text{if } \mathbb{E}_{\pi^2_{n(\theta,i)}}[I_1] M \geq 1, \\ 0 & \text{else.} \end{cases}$$

$$\text{For all } i \text{ in } I \text{ such that } \theta_i = 1,\ s^{2*}_{n(\theta,i)} = 0$$

Team reasoners do their bit if and only if the expected team size is large enough, and strategic reasoners shirk. Since, along the equilibrium path, second period beliefs are Bayesian updates and, for all i in I, $\lambda_i = \lambda$, assumption 3 (no mistakes in performance) guaranties that:

$$\text{For all } i, j \text{ in } I,\ j \neq i, \quad \pi_i[\cdot \mid \theta_i, a^1] = \pi_i[\cdot \mid \theta_i, s^1]$$

Thus, we just have to analyze first period prescriptions.

(i) An equilibrium without any contributions exists if and only if none of the players, strategic or team, has a reason to deviate. Note that when noone contributes the team player may have two reasons to deviate. A *direct* reason: given a high enough common prior, the team has a higher expected payoff from investing in the public good. An *indirect* reason: even in cases where the prior belief is not high enough to directly motivate a first

round contribution, prescribing team members to contribute in the first period yields —as a deviation from the no contributing equilibrium— perfect information about the team size for the second period. Under perfect information, an optimal second period choice can be made with certainty. Thus, the team player deviates if and only if:

$$\begin{aligned} 2 \;<\; & \mathbb{E}_\pi[I_1 \mid \theta_i = 0]M + \mathbb{P}_\pi(I_1 M \geq 1 \mid \theta_i = 0)\mathbb{E}_\pi[I_1 \mid \theta_i = 0, I_1 M \geq 1]M \\ & + \mathbb{P}_\pi(I_1 M < 1 \mid \theta_i = 0) \end{aligned}$$

That is,

$$\begin{aligned} 2 - \Big(2 + (I-1)\pi + \tilde{\mathbb{E}}_\pi[\mathbb{1}_{\{I_1 M \geq 1-M\}} I_1]\Big) M & \\ < \;\; (1-M)\tilde{\mathbb{P}}_\pi(I_1 M < 1 - M) & \end{aligned} \tag{C.1}$$

where $\tilde{\mathbb{P}}_\pi(\cdot)$ and $\tilde{\mathbb{E}}_\pi(\cdot)$ result from considering trials among $I-1$, not I, individuals: all individuals but individual i.[1] Both the left hand side and the right hand side are continuously decreasing fucntions over $[0,1]$. Lemma C.1 establishes that, in the presence of 4 individuals, for any admissible values of M, the two curves cross only once. Denote π^l_M the value of π where the two curves cross. It is well defined for any M above $1/4$.

A strategic player deviates if and only if his deviation brings about higher chances of contributions by a sufficiently large team in period 2. No combinations of π and M can be such that strategic reasoners have a reason to contribute in the first period when team reasoners do not. Indeed, considering the most optimistic off-path belief possible—one that takes the deviating contribution to constitute doubtless evidence about the team reasoner status of the contributor, we obtain that the deviation will trigger a second period contribution by team reasoners if and only if

$$\mathbb{E}_\pi[I_1 \mid \theta_i = 0] = 1 + (I-1)\pi \in \left[\tfrac{1}{M} - 1, \tfrac{1}{M}\right] \;\; \Leftrightarrow \;\; \pi \in \left[\tfrac{1-2M}{(I-1)M}, \tfrac{1-M}{(I-1)M}\right]$$

But, for such values of π, the expected team size is too small to make deviations attractive to strategic reasoners, since

$$2 < M + 1 + \mathbb{E}_\pi[I_1 \mid \theta_i = 1]M \;\; \Leftrightarrow \;\; \pi > \tfrac{1-M}{(I-1)M}$$

This establishes the first statement.

(ii) Any equilibrium in which the team contributes in the first period when strategic players do not is characterized by reasons to deviate. Equation (C.1) in the proof of (i) guarantees the existence, for each M, of a unique threshold, π^l_M, below which the team

[1] The reformulation, permitted by our assumption of independence between types, is useful because it avoids conditioning on a 0-probability event whenever $\pi = 0$.

has a reason to deviate from the equilibrium where only the team contributes and after which it has none. Assume π is larger than π_M^l and consider the incentives of a strategic player. He has a reason to mimic team reasoners whenever his deviation brings about higher chances of contribution by a sufficiently large team in period 2. This will happen in exactly one event: when the realized team size is $\lfloor \frac{1}{M} \rfloor$. Thus, imitating team-reasoners is profitable in expectation if and only if:

$$1 - M \;<\; \mathbb{P}_\pi(I_1 = \lfloor \tfrac{1}{M} \rfloor \mid \theta_i = 1) \lfloor \tfrac{1}{M} \rfloor M$$

That is, if and only if $M > \frac{1}{4}$ and

$$1 \;<\; M + \binom{I-1}{\lfloor \frac{1}{M} \rfloor} \pi^{\lfloor \frac{1}{M} \rfloor} (1-\pi)^{I-1-\lfloor \frac{1}{M} \rfloor} \lfloor \tfrac{1}{M} \rfloor M \tag{C.2}$$

When M belongs to $(1/4, 1/3]$, the right hand side of equation (C.2) takes value $M(1 + 3\pi^3)$. Therefore, equation (C.2) holds if and only if

$$\pi > \left(\tfrac{1-M}{3M}\right)^{1/3}$$

For any such M, denote $\pi_M^m \equiv \max\{\pi_M^l, ((1-M)/3M)^{1/3}\}$. This threshold is uniquely defined and smaller than $1 =: \pi_M^h$.

Consider now the case where $M \in (\frac{1}{3}, \frac{1}{2}]$. The right hand side of equation (C.2) takes value $M(1 + 6\pi^2(1-\pi))$, so that equation (C.2) holds if and only if

$$\pi^2(1-\pi) > \tfrac{1-M}{6M}$$

Since the maximum of the left hand side over $[0, 1]$ is smaller than the minimum of the right hand side over $(\frac{1}{3}, \frac{1}{2}]$, equation (C.2) never holds for such a M. Define $\pi_M^m \equiv \pi_M^h \equiv 1$

The case where $(\frac{1}{2}, 1]$ is more subtle. Note first that, whenever M belongs to this interval, the right hand side of equation (C.2) takes value $M(1 + 3\pi(1-\pi)^2)$. Hence, it holds if and only if

$$\pi(1-\pi)^2 > \frac{1-M}{3M} \tag{C.3}$$

The left hand side reaches a maximum for $\pi^{max} = 1/3$ and the right hand side is decreasing in M. Thus, a threshold M can be derived below which equation (C.2) never holds. Namely,

$$M = \tfrac{1}{1+3\pi^{max}(1-\pi^{max})^2} = 9/13$$

This settles the case for M in $(\frac{1}{2}, \frac{9}{13}]$. When M belongs to $(9/13, 1]$, the right-hand side in equation (C.3) is strictly lower than the maximum of the left hand side. $\pi(1-\pi)^2$ being

bell shaped, the constant map $(1-M)/3M$ intersect it twice. Call π_M^m the maximum of π_M^l and the abscissa of the first intersection, and π_M^h that of the second. Within that interval, equation (C.2) holds. Outside, it does not.

(iii) (Proof for any $I \geq 3$) Assume, for instance, that an equilibrium could be sustained which involves participation by all reasoning types in the first period. Then we must have $\pi \geq \frac{1-M}{(I-1)M}$ since otherwise the following equation

$$1 - M \;\; > \;\; \mathbb{E}_\pi[I_1 \mid \theta_i = 1]M$$

would hold, which says that, for strategic reasoners, benefits from separation are larger than expected benefits from pooling. Furthermore, the absence of a contribution by a strategic player i leads to non-contribution by team reasoners in the second period if and only if team reasoner's off-path beliefs are such that:

$$\mathbb{E}_\pi[I_1 \mid \theta_i = 0, a_{-i}^1 = 1, a_i^1 = 0] < \frac{1}{M} \tag{C.4}$$

Such off-path beliefs can be specified if and only if $\pi \leq \frac{1-M}{(I-2)M}$. Thus, for all M in $[1/4, 1]$, the existence of equilibria involving full contribution in the first period entail:

$$\frac{1-M}{(I-1)M} \leq \pi \leq \min\left\{1, \frac{1-M}{(I-2)M}\right\}$$

■

C.2 Octave Code for Figures 1-3

C.2.1 Figure 1

```
clear
% Equilibria for I=4
m=0.00:0.001:1; % possible values of the transformation rate
p=0.00:0.001:1; % possible values of the prior belief

z=zeros(1001,1001);

y=zeros(1001,1000);

for i = 1:1001
```

```
for j = 501:1001
if ((1-p(i))^3*(1-m(j))<2-2*m(j)-6*p(i)*m(j))
y(i,j)=-1; % Opposite of equation (C.1)
else
y(i,j)=m(j)*(1+nchoosek(3,floor(1/m(j)))*p(i)^(floor(1/m(j)))
    *(1-p(i))^(3-floor(1/m(j)))*floor(1/m(j))); % Right hand side
     of equation (C.2)
if ((1-m(j))/(3*m(j))<=p(i))
if ((1-m(j))/(2*m(j))>=p(i))
z(i,j)=2; % Condition (C.4)
end
end
end
end
end

for i = 1:1001
for j = 335:500
if ((1+2*p(i)^3-3*p(i)^2)*(1-m(j))<2-2*m(j)-(3*p(i)+6*p(i)^2-3*p
    (i)^3)*m(j))
y(i,j)=-1; % Opposite of equation (C.1)
else
y(i,j)=m(j)*(1+nchoosek(3,floor(1/m(j)))*p(i)^(floor(1/m(j)))
    *(1-p(i))^(3-floor(1/m(j)))*floor(1/m(j))); % Right hand side
     of equation (C.2)
if ((1-m(j))/(3*m(j))<=p(i))
if ((1-m(j))/(2*m(j))>=p(i))
z(i,j)=2; % Condition (C.4)
end
end
end
end
end

for i = 1:1001
for j = 252:334
if ((1-p(i)^3)*(1-m(j))<2-2*m(j)-(3*p(i)+3*p(i)^3)*m(j))
```

```
43 y(i,j)=-1; % Opposite of equation (C.1)
44 else
45 y(i,j)=m(j)*(1+nchoosek(3,floor(1/m(j)))*p(i)^(floor(1/m(j)))
       *(1-p(i))^(3-floor(1/m(j)))*floor(1/m(j))); % Right hand side
        of equation (C.2)
46 if ((1-m(j))/(3*m(j))<=p(i))
47 if ((1-m(j))/(2*m(j))>=p(i))
48 z(i,j)=2; % Condition (C.4)
49 end
50 end
51 end
52 end
53 end
54
55 for i = 1:1001
56 for j = 1:251
57 y(i,j)=-1;
58 end
59 end
60
61 %[mm,pp] = meshgrid(m,p);
62
63 colormap ("winter");
64
65 contour (p,m,y,[-1,-1]:[1 1])
66 hold on
67 contour (p,m,z,[-1,-1]:[2 2])
68 hold off
69 title ({"Case I=4"});
```

C.2.2 Figure 2

```
1 clear
2 % Equilibria for I=10
3 m=0.00:0.001:1; % possible values of the transformation rate
4 p=0.00:0.001:1; % possible values of the prior belief
5
6 w=zeros(1001,9);
```

```
x=zeros(1001,9);
z=zeros(1001,1001);

for i=1:1001
for j= 1:9
x(i,j)= nchoosek(9,j)*p(i)^j*(1-p(i))^(9-j); % Basis for
    equation (C.1)
w(i,j)= nchoosek(9,j)*j*p(i)^j*(1-p(i))^(9-j); % Basis for
    equation (C.1)
end
end

y=zeros(1001,1000);

for i = 1:1001
for j = 501:1001
if ((1-p(i))^9*(1-m(j))<2-2*m(j)-18*p(i)*m(j))
y(i,j)=-1; % Opposite of equation (C.1)
else
y(i,j)=m(j)*(1+nchoosek(9,floor(1/m(j)))*p(i)^(floor(1/m(j)))
    *(1-p(i))^(9-floor(1/m(j)))*floor(1/m(j))); % Right hand side
     of equation (C.2)
if ((1-m(j))/(9*m(j))<=p(i))
if ((1-m(j))/(8*m(j))>=p(i))
z(i,j)=2; % Condition (C.4)
end
end
end
end
end

for i = 1:1001
for j = 335:500
if (((1-p(i))^9+x(i,1))*(1-m(j))<2-2*m(j)-(18*p(i)-w(i,1))*m(j))
y(i,j)=-1; % Opposite of equation (C.1)
else
y(i,j)=m(j)*(1+nchoosek(9,floor(1/m(j)))*p(i)^(floor(1/m(j)))
```

```
    *(1-p(i))^(9-floor(1/m(j)))*floor(1/m(j)); % Right hand side
     of equation (C.2)
40 if ((1-m(j))/(9*m(j))<=p(i))
41 if ((1-m(j))/(8*m(j))>=p(i))
42 z(i,j)=2; % Condition (C.4)
43 end
44 end
45 end
46 end
47 end
48
49 for i = 1:1001
50 for j = 102:334
51 if (((1-p(i))^9+sum(x(i,[1:ceil(1/m(j)-2)])))*(1-m(j))<2-2*m(j)
    -(18*p(i)-sum(w(i,[1:ceil(1/m(j)-2)])))*m(j))
52 y(i,j)=-1; % Opposite of equation (C.1)
53 else
54 y(i,j)=m(j)*(1+nchoosek(9,floor(1/m(j)))*p(i)^(floor(1/m(j)))
    *(1-p(i))^(9-floor(1/m(j)))*floor(1/m(j))); % Right hand side
     of equation (C.2)
55 if ((1-m(j))/(9*m(j))<=p(i))
56 if min(1,((1-m(j))/(8*m(j)))>=p(i))
57 z(i,j)=2; % Condition (C.4)
58 end
59 end
60 end
61 end
62 end
63
64 for i = 1:1001
65          for j = 1:101
66          y(i,j)=-1;
67          end
68 end
69
70 %[mm,pp] = meshgrid(m,p);
71
```

```
72 colormap ("winter");
73
74 contour (p,m,y,[-1,-1]:[1 1])
75 hold on
76 contour (p,m,z,[-1,-1]:[2 2])
77 hold off
78 title ({"Case I=10"});
```

C.2.3 Figure 3

```
1 clear
2 % Equilibria for I=25
3 m=0.00:0.001:1; % possible values of the transformation rate
4 p=0.00:0.001:1; % possible values of the prior belief
5
6 w=zeros(1001,24);
7 x=zeros(1001,24);
8 z=zeros(1001,1001);
9
10 for i=1:1001
11 for j= 1:24
12 x(i,j)= nchoosek(24,j)*p(i)^j*(1-p(i))^(24-j); % Basis for
     equation (C.1)
13 w(i,j)= nchoosek(24,j)*j*p(i)^j*(1-p(i))^(24-j); % Basis for
     equation (C.1)
14 end
15 end
16
17 y=zeros(1001,1000);
18
19 for i = 1:1001
20 for j = 501:1001
21 if ((1-p(i))^24*(1-m(j))<2-2*m(j)-48*p(i)*m(j))
22 y(i,j)=-1; % Opposite of equation (C.1)
23 else
24 y(i,j)=m(j)*(1+nchoosek(24,floor(1/m(j)))*p(i)^(floor(1/m(j)))
     *(1-p(i))^(24-floor(1/m(j)))*floor(1/m(j))); % Right hand
     side of equation (C.2)
```

```
25 if ((1-m(j))/(24*m(j))<=p(i))
26 if ((1-m(j))/(23*m(j))>=p(i))
27 z(i,j)=2; % Condition (C.4)
28 end
29 end
30 end
31 end
32 end
33
34 for i = 1:1001
35 for j = 335:500
36 if (((1-p(i))^24+x(i,1))*(1-m(j))<2-2*m(j)-(48*p(i)-w(i,1))*m(j)
   )
37 y(i,j)=-1; % Opposite of equation (C.1)
38 else
39 y(i,j)=m(j)*(1+nchoosek(24,floor(1/m(j)))*p(i)^(floor(1/m(j)))
   *(1-p(i))^(24-floor(1/m(j)))*floor(1/m(j))); % Right hand
   side of equation (C.2)
40 if ((1-m(j))/(24*m(j))<=p(i))
41 if ((1-m(j))/(23*m(j))>=p(i))
42 z(i,j)=2; % Condition (C.4)
43 end
44 end
45 end
46 end
47 end
48
49 for i = 1:1001
50 for j = 42:334
51 if (((1-p(i))^24+sum(x(i,[1:ceil(1/m(j)-2)])))*(1-m(j))<2-2*m(j)
   -(48*p(i)-sum(w(i,[1:ceil(1/m(j)-2)])))*m(j))
52 y(i,j)=-1; % Opposite of equation (C.1)
53 else
54 y(i,j)=m(j)*(1+nchoosek(24,floor(1/m(j)))*p(i)^(floor(1/m(j)))
   *(1-p(i))^(24-floor(1/m(j)))*floor(1/m(j))); % Right hand
   side of equation (C.2)
55 if ((1-m(j))/(24*m(j))<=p(i))
```

```
56 if min(1,((1-m(j))/(23*m(j)))>=p(i))
57 z(i,j)=2; % Condition (C.4)
58 end
59 end
60 end
61 end
62 end
63
64 for i = 1:1001
65           for j = 1:41
66           y(i,j)=-1;
67           end
68 end
69
70 %[mm,pp] = meshgrid(m,p);
71
72 colormap ("winter");
73
74 %subplot(1,3,2)
75 contour (p,m,y,[-1,-1]:[1 1])
76 hold on
77 contour (p,m,z,[-1,-1]:[2 2])
78 hold off
79 title ({"Case I=25"});
```

Bibliography

Ambrus, A. and P. A. Pathak

2011. Cooperation over finite horizons: A theory and experiments. *Journal of Public Economics*, 95(7):500–512.

Andreoni, J.

1988. Privately provided public goods in a large economy: the limits of altruism. *Journal of Public Economics*, 35(1):57–73.

Andreoni, J.

1989. Giving with impure altruism: Applications to charity and ricardian equivalence. *Journal of Political Economy*, 97(6):1447–1458.

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. and R. Croson

2008. Partners versus strangers: Random rematching in public goods experiments. *Handbook of experimental economics results*, 1:776–783.

Andreoni, J., J. M. Rao, and H. Trachtman

2017. Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy*, 125(3):625–653.

Andreoni, J. and L. Samuelson

2006. Building rational cooperation. *Journal of Economic Theory*, 127(1):117–154.

Anscombe, G. E. M.

1958. On brute facts. *Analysis*, 18(3):69–72.

Anscombe, G. E. M.

1981 [1975]. The first person. In *The Collected Philosophical Papers of G.E.M. Anscombe*, volume 96. Basil Blackwell - Oxford.

Anscombe, G. E. M.
2000 [1957]. *Intention*. Harvard University Press.

Anscombe, G. E. M.
2001 [1971]. *An introduction to Wittgenstein's Tractatus*. St. Augustine's Press, South Bend, India.

Arendt, H.
1972. Lying in politics. In *Crises of the Republic*, Pp. 3–47. Harcourt Brace New York.

Arifovic, J. and J. Ledyard
2012. Individual evolutionary learning, other-regarding preferences, and the voluntary contributions mechanism. *Journal of Public Economics*, 96(9):808–823.

Arrow, K. J.
2012 [1951, 1963]. *Social choice and individual values*, volume 12. Yale university press.

Aumann, R. J.
1987. Correlated equilibrium as an expression of bayesian rationality. *Econometrica*, 55(1):1–18.

Aumann, R. J.
1998. Common priors: A reply to g$'$ul. *Econometrica*, 66(4):929–938.

Aumann, R. J.
2000. *Collected Papers*. MIT Press.

Austin, J. L.
1979. Performative utterances. In *Philosophical Papers*, J. O. Urmson and G. J. Warnock, eds., Pp. 232–252. Clarendon Press.

Ayer, A. J.
2014 [1946]. *Language, truth and logic (2nd edition)*. Dover Publications, Inc., New York.

Bacharach, M.
2006. *Beyond individual choice: teams and frames in game theory*. Princeton University Press.

Bachelard, G.
2011 [1938]. *La formation de l'esprit scientifique: contribution à une psychanalyse de la connaissance*. Vrin. The Formation of the Scientific Mind. Clinamen, Bolton, 2002.

Bénabou, R. and J. Tirole
2006. Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.

Bénabou, R. and J. Tirole
2011. Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, 126(2):805–855.

Bergstrom, T., L. Blume, and H. Varian
1986. On the private provision of public goods. *Journal of Public Economics*, 29(1):25–49.

Böll, H.
2017 [1963]. *Ansichten eines Clowns* [*The Clown*]. Kiepenheuer & Witsch.

Bolton, G. E. and A. Ockenfels
2000. ERC: A theory of equity, reciprocity, and competition. *American Economic review*, Pp. 166–193.

Buchanan, J. M.
1962. The relevance of pareto optimality. *Journal of conflict resolution*, 6(4):341–354.

Buckley, E. and R. Croson
2006. Income and wealth heterogeneity in the voluntary provision of linear public goods. *Journal of Public Economics*, 90(4):935–955.

Camus, A.
1971 [1942]. L'étranger [the stranger]. *Paris, Gallimard*.

Canguilhem, G.
1955. *La formation du concept de réflexe aux XVIIe et XVIIIe siècles*. Presses Universitaires de France.

Carroll, L.
1895. What the tortoise said to achilles. *Mind*, 4(14):278–280.

Cavaillès, J.
1935. L'école de vienne au congrès de prague. *Revue de Métaphysique et de Morale*, 42(1):137–149.

Chakravartty, A.
2017. Scientific realism. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, ed. Metaphysics Research Lab, Stanford University.

Charness, G. and M. Rabin
2002. Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869.

Chaudhuri, A.

2011. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, 14(1):47–83.

Coase, R. H.

1982. How should economists choose? *The G. Warren Nutter Lectures in Political Economy*.

Cornes, R. and R. Hartley

2007. Aggregative public good games. *Journal of Public Economic Theory*, 9(2):201–219.

Crumpler, H. and P. J. Grossman

2008. An experimental test of warm glow giving. *Journal of Public Economics*, 92(5-6):1011–1021.

Daston, L. and P. Galison

2007. *Objectivity*. Zone books.

Descartes, R.

2008 [1637]. *Discourse on the Method*. Oxford University Press.

Dietrich, F. and C. List

2016. Mentalism versus behaviourism in economics: a philosophy-of-science perspective. *Economics and Philosophy*, 32(02):249–281.

Elster, J.

2009. *Le désintéressement. Traité critique de l'homme économique. Tome I.* Éditions du Seuil.

Elster, J.

2011. The valmont effect: The warm-glow theory of philanthropy. *Giving well: The ethics of philanthropy*, Pp. 67–83.

Falk, A. and U. Fischbacher

2006. A theory of reciprocity. *Games and Economic Behavior*, 54(2):293–315.

Fehr, E. and S. Gächter

2000. Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14(3):159–181.

Fehr, E. and K. M. Schmidt

1999. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Ferguson, E., M. Taylor, D. Keatley, N. Flynn, and C. Lawrence
2012. Blood donors' helping behavior is driven by warm glow: more evidence for the blood donor benevolence hypothesis. *Transfusion*, 52(10):2189–2200.

Foot, P.
2001. *Natural goodness*. Oxford University Press.

Frege, G.
1948 [1892]. Sense and reference. *The philosophical review*, 57(3):209–230.

Friedman, M.
2008 [1953]. The methodology of positive economics. In *The Philosophy of Economics - An Anthology*, D. M. Hausman, ed., Pp. 145–178. Cambridge University Press.

Fudenberg, D. and D. K. Levine
1998. *The theory of learning in games*, volume 2. MIT press.

Gary, R.
2013 [1956]. *Les racines du ciel*. Editions Gallimard.

Gilbert, M.
1990. Walking together: A paradigmatic social phenomenon. *MidWest studies in philosophy*, 15(1):1–14.

Gilbert, M.
1992 [1989]. *On social facts*. Princeton University Press.

Gilbert, M.
2015. Joint commitment: What it is and why it matters. *Phenomenology and Mind*, 9:18–26.

Goffman, E.
1990 [1959]. *The presentation of self in everyday life*. Cox & Wyman Ltd., Reading.

Gold, N. and R. Sugden
2007. Collective intentions and team agency. *The Journal of Philosophy*, 104(3):109–137.

Gul, F. and W. Pesendorfer
2008. The case for mindless economics. In *The Foundations of Positive and Normative Economics: A handbook*, A. Caplin and A. Schotter, eds., Pp. 3–42. Oxford University Press.

Harsanyi, J. C.
1977a. Morality and the theory of rational behavior. *Social research*, Pp. 623–656.

Harsanyi, J. C.
1977b. Rule utilitarianism and decision theory. *Erkenntnis*, 11(1):25–53.

Harsanyi, J. C.
1982. Rule utilitarianism, rights, obligations and the theory of rational behavior. In *Papers in Game Theory*, Pp. 235–253. Springer.

Hart, S.
2011. Commentary: Nash equilibrium and dynamics. *Games and Economic Behavior*, 71(1):6–8.

Hart, S. and Y. Mansour
2010. How long to equilibrium? the communication complexity of uncoupled equilibrium procedures. *Games and Economic Behavior*, 69(1):107–126.

Hart, S. and A. Mas-Colell
2003. Uncoupled dynamics do not lead to nash equilibrium. *The American Economic Review*, 93(5):1830–1836.

Hart, S. and A. Mas-Colell
2013. *Simple adaptive strategies: from regret-matching to uncoupled dynamics*, volume 4. World Scientific.

Hausman, D. M.
1994. Why look under the hood? *The Philosophy of Economics: An Anthology*, Pp. 217–221.

Hausman, D. M.
2011. Mistakes about preferences in the social sciences. *Philosophy of the social sciences*, 41(1):3–25.

Hayek, F. A. v.
1937. Economics and knowledge. *Economica*, 4(13):33–54.

Hayek, F. A. v.
1975. The pretence of knowledge. *The Swedish Journal of Economics*, 77(4):433–442.

Hempel, C. G.
1980 [1935]. The logical analysis of psychology. In *Readings in philosophy of psychology*, N. J. Block, ed., volume 1, Pp. 14–23. Harvard University Press Cambridge, MA.

Holton, G.
1995. On the vienna circle in exile: an eyewitness report. In *The Foundational Debate*, Pp. 269–292. Springer.

Hume, D.
1907 [1741]. *Essays: Moral, Political, and Literary*, volume 1. Longmans, Green, and Company.

Hume, D.
1969 [1739]. *A treatise of human nature*. Penguin Classics.

Kolm, S.-C.
1969 [1965]. The optimal production of social justice. In *Public Economics: An Analysis of Public Production and Consumption and their Relations to the Private Sectors*, J. Margolis and H. Guitton, eds., Pp. 145–200. London: Macmillan.

Kreps, D. M., P. Milgrom, J. Roberts, and R. Wilson
1982. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic theory*, 27(2):245–252.

Kuhn, T. S.
2012 [1962]. *The structure of scientific revolutions*. University of Chicago press.

Laffont, J.-J.
1975. Macroeconomic constraints, economic efficiency and ethics: An introduction to kantian economics. *Economica*, 42(168):430–437.

Laury, S. K. and C. A. Holt
2008. Voluntary provision of public goods: experimental results with interior nash equilibria. *Handbook of Experimental Economics Results*, 1:792–801.

Ledyard, J. O.
1995. Public goods: A survey of experiemental research. In *The Handbook of Experimental Economics, Volume 1*, J. H. Kagel and A. E. Roth, eds., Pp. 111–194. Princeton university press.

Luce, R. D. and H. Raiffa
2012 [1957]. *Games and decisions: Introduction and critical survey*. Courier Corporation.

Maddy, P.
2012. The philosophy of logic. *Bulletin of Symbolic Logic*, 18(4):481–504.

Mäki, U.
2006. On the method of isolation in economics. In *Recent Developments in Economic Methodology, Vol. 3*, J. B. Davis, ed. Edward Edgar.

Malraux, A.
1972 [1933]. La condition humaine [man's fate]. *Paris, Gallimard*.

McCloskey, D. N.
1983. The rhetoric of economics. *Journal of Economic Literature*, 21(2):481–517.

McGeer, V.
2007. The regulative dimension of folk psychology. In *Folk psychology re-assessed*, Pp. 137–156. Springer.

Morgenstern, O.
1976. The collaboration between oskar morgenstern and john von neumann on the theory of games. *Journal of Economic Literature*, 14(3):805–816.

Okasha, S.
2016. On the interpretation of decision theory. *Economics and Philosophy*, Pp. 1–25.

Ostrom, E.
2015 [1990]. *Governing the commons*. Cambridge university press.

Palfrey, T. R. and J. E. Prisbrey
1997. Anomalous behavior in public goods experiments: How much and why? *The American Economic Review*, Pp. 829–846.

Plato
-380. Book 7. In *Republic*.

Plato
-385. Crito. In *Plato in Twelve Volumes.*, volume 1.

Polanyi, M.
2015 [1958]. *Personal knowledge: Towards a post-critical philosophy*. University of Chicago Press.

Quine, W. V.
1948. On what there is. *The Review of Metaphysics*, 2(1):21–38.

Quine, W. V.
1951. Main trends in recent philosophy: Two dogmas of empiricism. *The philosophical review*, Pp. 20–43.

Ramsey, F. P.
1931. Truth and probability (1926). *The foundations of mathematics and other logical essays*, Pp. 156–198.

Ravenscroft, I.
2019. Folk psychology as a theory. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, ed. Metaphysics Research Lab, Stanford University.

Rawls, J.
1955. Two concepts of rules. *The philosophical review*, 64(1):3–32.

Ribar, D. C. and M. O. Wilhelm
2002. Altruistic and joy-of-giving motivations in charitable behavior. *Journal of Political Economy*, 110(2):425–457.

Roemer, J. E.
2010. Kantian equilibrium. *The Scandinavian Journal of Economics*, 112(1):1–24.

Roemer, J. E.
2015. Kantian optimization: A microfoundation for cooperation. *Journal of Public Economics*, 127:45–57.

Rousseau, J.-J.
1997 [1755]. A discourse on political economy. In *The Social Contract and Other Later Political Writings*. Cambridge University Press.

Rousseau, J.-J.
1997 [1762]. The social contract. In *The Social Contract and Other Later Political Writings*. Cambridge University Press.

Rousseau, J.-J.
2011 [1782]. *The reveries of the solitary walker*. Oxford University Press.

Rubinstein, A.
1991. Comments on the interpretation of game theory. *Econometrica*, Pp. 909–924.

Rubinstein, A.
2012. *Economic fables*. Open book publishers.

Rubinstein, A. and Y. Salant
2008. Some thoughts on the principle of revealed preference. In *The Foundations of Positive and Normative Economics: A Handbook*, A. Caplin and A. Schotter, eds., Pp. 115–124. Oxford University Press.

Runciman, W. and A. Sen
1965. Games, justice and the general will. *Mind*, 74(296):554–562.

Russell, B.
1905. On denoting. *Mind*, 14(56):479–493.

Russell, B.
2001 [1912]. *The problems of philosophy*. Oxford University Press.

Russell, B.
2010 [1919]. Descriptions. In *Introduction to Mathematical Philosophy*, Pp. 90–100. George Allen & Unwin, Ltd., London.

Samuelson, P. A.
1938. A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61–71.

Samuelson, P. A.
1948. *Foundations of economic analysis*. Harvard University Press.

Samuelson, P. A.
1963. Discussion: problems of methodology. In *American Economic Review, Proceedings*, volume 53, Pp. 231–236.

Sartre, J.-P.
2000 [1938]. *La nausée* [*Nausea*]. Paris: Gallimard.

Savage, L. J.
1972. *The foundations of statistics*. Courier Corporation.

Scheer, R.
2004. The 'mental state'theory of intentions. *Philosophy*, 79(1):121–131.

Schlick, M.
1982 [1934]. The foundations of knowledge. In *Logical Positivism*, A. J. Ayer, ed. The Free Press, New York.

Schmid, H. B.
2016. On knowing what we're doing together: Groundless group self-knowledge and plural self-blindness. In *The epistemic life of groups: essays in the epistemology of collectives*, Pp. 51–74. Oxford University Press.

Schmid, H. B.
2018. The subject of "we intend". *Phenomenology and the Cognitive Sciences*, 17(2):231–243.

Sellars, W.

1956. Empiricism and the philosophy of mind. *Minnesota studies in the philosophy of science*, 1(19):253–329.

Sen, A.

1973. Behaviour and the concept of preference. *Economica*, 40(159):241–259.

Sen, A.

1974. Choice, ordering and morality. In *Practical Reason*, S. Körner, ed. Oxford: Blackwell.

Sen, A.

1985. Rationality and uncertainty. *Theory and Decision*, 18.

Sen, A.

1993. Internal consistency of choice. *Econometrica*, Pp. 495–521.

Sen, A.

1997. Maximization and the act of choice. *Econometrica*, Pp. 745–779.

Sen, A. K.

1977. Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs*, Pp. 317–344.

Seneca

1900 [58 AD]. Of a happy life. In *L. Annaeus Seneca, Minor Dialogs Together with the Dialog "On Clemency"*, T. S. Aubrey, ed. London: George Bell and Sons.

Smith, A.

2003 [1776]. *An Inquiry into the Nature and Causes of the Wealth of Nations. Books I-III*. Penguin Classics.

Spiller, J., A. Ufert, P. Vetter, and U. Will

2016. Norms in an asymmetric public good experiment. *Economics Letters*, 142:35–44.

Sugden, R.

1982. On the economics of philanthropy. *The Economic Journal*, 92(366):341–350.

Sugden, R.

1984. Reciprocity: the supply of public goods through voluntary contributions. *The Economic Journal*, 94(376):772–787.

Sugden, R.

1993. Thinking as a team: Towards an explanation of nonselfish behavior. *Social philosophy and policy*, 10(01):69–89.

Sugden, R.

2000. Credible worlds: the status of theoretical models in economics. *Journal of Economic Methodology*, 7(1):1–31.

Tirole, J.

1999. Incomplete contracts: Where do we stand? *Econometrica*, 67(4):741–781.

Warr, P. G.

1982. Pareto optimal redistribution and private charity. *Journal of Public Economics*, 19(1):131–138.

Weber, M.

1949 [1904]. " objectivity" in social science and social policy. *The methodology of the social sciences*, Pp. 49–112.

Wittgenstein, L.

1961 [1921]. *Tractatus-Logico-Philosophicus*. Humanities Press. New-York. transl. D. F. Pears and B. F. McGuinness.

Wittgenstein, L.

2009 [1953]. *Philosophical Investigations* (*Philosophische Untersuchungen.*) *Eng. & Ger*. Wiley-Blackwell. G. E. M. Anscombe's translation.

Curriculum Vitae

2007–2010 B.Sc. in Economics, University of Limoges

2010–2012 M.Sc. in Economics, Toulouse School of Economics

2012–2019 Ph.D. in Economics, University of Mannheim

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbstständig angefertigt und die benutzten Hilfsmittel vollständig und deutlich angegeben habe.

Mannheim, 24.07.2019

Justin Leduc