# Reconciling Normative and Behavioural Economics

Guilhem Lecouteux

**ÉCOLE DOCTORALE DE L'ÉCOLE POLYTECHNIQUE**

# T H È S E

pour l'obtention du titre de

**Docteur de l'École Polytechnique**

**Mention : Sciences Économiques**

Présentée et soutenue par

Guilhem Lecouteux

# Reconciling Normative and Behavioural Economics

Thèse dirigée par Francis Bloch et Robert Sugden

soutenue le 27 Mai 2015

**Jury :**

| | | | |
|---|---|---|---|
| *Rapporteurs* : | Franz Dietrich | - | CNRS & Paris School of Economics |
| | Natalie Gold | - | King's College London |
| *Directeurs* : | Francis Bloch | - | Paris School of Economics |
| | Robert Sugden | - | University of East Anglia |
| *Président* : | Richard Arena | - | Université Nice - Sophia Antipolis |
| *Examinateurs* : | Till Grüne-Yanoff | - | KTH Royal Institute of Technology |
| | Guillaume Hollard | - | CNRS & École Polytechnique |
| | Marco Mariotti | - | Queen Mary University of London |

# Remerciements

Rédiger mes remerciements de thèse constitue un moment privilégié dans mon parcours de chercheur, car cela signifie la fin d'une importante étape, mais aussi que le moment est venu de remercier comme il se doit toutes les personnes qui m'ont supporté et sans lesquelles cette thèse n'aurait pas été ce qu'elle est aujourd'hui. C'est également une tâche difficile, car — contrairement au reste de ce manuscrit qui n'intéressera probablement que les lecteurs intéressés — il est fort probable que, parmi les rares personnes qui tiendront un jour entre leurs mains ce manuscrit, tous finiront nécessairement par lire ces quelques lignes (un des principaux enseignements que j'ai retiré de mon parcours universitaire est en effet que, quelque soit la discipline, toute personne lisant une thèse est irrésistiblement attirée par la section contenant les remerciements de son auteur). Je ferai donc de mon mieux pour mener cette tâche à bien, et je m'excuse par avance auprès de toutes les personnes que j'aurais omises de citer, ou qui s'estimeraient lésées de ne pas avoir la place qu'elles méritent (afin d'éviter tout soupçon de favoritisme, les noms sont donnés par ordre chronologique puis alphabétique).

Mes premiers remerciements s'adressent tout naturellement à mes deux directeurs de thèse, Francis Bloch et Bob Sugden. Ce travail leur doit énormément et j'ai pleinement conscience de l'immense chance que j'ai eue d'avoir pu effectuer ma thèse sous leur direction. J'ai rencontré Francis en 2011 au cours de mon master, et j'envisageais alors de commencer une thèse en théorie des jeux sur les négociations internationales. Ayant déjà trouvé un stage, mais disposant également d'une année de financement supplémentaire de la part de l'ENS Cachan entre mon master et la thèse, nous avons décidé que j'effectuerai un stage de recherche d'un an avant de débuter officiellement ma thèse. La réalisation de mon mémoire de master m'a néanmoins fait bifurquer vers des thématiques plus générales liées au choix rationnel et aux problèmes d'action collective, et Francis a accepté de m'encadrer en thèse à partir de Septembre 2012 sur les aspects de théorie des jeux de mon projet. Souhaitant également travailler sur des aspects plus philosophiques du choix rationnel, Francis a spontanément évoqué Bob comme un possible co-encadrant de thèse, qui a lui aussi accepté de superviser ma recherche. Bob m'a alors accueilli durant ma première année de thèse à Norwich, à une période où mon projet de recherche était encore en construction. C'est durant ce séjour que j'ai découvert le problème de réconciliation, qui constitue désormais la thématique centrale de cette thèse, auquel j'ai pu appliquer le modèle de préférences sur lequel je travaillais depuis mon stage de master. Je les remercie

tous deux pour avoir su m'offrir un encadrement d'une excellente qualité — je me suis assez rapidement rendu compte que beaucoup de thésards n'avaient pas ma chance d'avoir des directeurs prêts à rencontrer leurs étudiants sur une base hebdomadaire... —, pour la grande liberté qu'ils m'ont laissée dans la définition de mon projet de recherche, et ainsi que pour leur patience et leur soutien sans faille, malgré ma légère hybris et mon caractère qui a pu se révéler obstiné de temps à autre. C'est aussi la complémentarité de deux directeurs qu'apparemment peu de choses réunissaient d'un point de vue académique qui a fait de cette thèse une expérience inoubliable sur le plan humain et intellectuel. Merci à vous deux pour tout.

Je tiens également à remercier Franz Dietrich et Natalie Gold pour avoir accepté d'être rapporteurs de cette thèse, ainsi que Richard Arena, Till Grüne-Yanoff, Guillaume Hollard et Marco Mariotti pour le temps qu'ils ont consacré à faire progresser ma recherche en acceptant de participer à mon jury de thèse. Il ne fait aucun doute que leurs commentaires et suggestions au cours de la présoutenance ont contribué à améliorer significativement la qualité de ce travail. Je les remercie par avance pour notre discussion au cours de la soutenance.

Cette thèse a aussi bénéficié de diverses collaborations que j'ai pu développer au cours de ces dernières années. Je tiens tout d'abord à adresser un remerciement tout particulièrement hétérodoxe à Léonard Moulin, co-auteur d'un article sur les potentiels effets ségrégatifs des frais d'inscriptions dans l'enseignement supérieur, qui a trouvé sa place dans sa propre thèse plutôt que dans celle-ci. Pour avoir trouvé un *nudge* adéquat il y a quelques années, il peut être considéré comme le principal responsable de mon addiction aux conférences, ce qui a indirectement contribué à profondément enrichir cette thèse, et directement contribué à épuiser les fonds de recherche de Francis. Je tiens également à remercier Gerardo Infante, co-auteur avec Bob du premier chapitre de cette thèse. Enfin, je remercie Lauren Larrouy, co-autrice d'un article qui se situe dans la directe continuation de la seconde partie de cette thèse, mais qui n'était pas encore suffisamment finalisé pour pouvoir être intégré dans le présent manuscrit.
En plus des chercheurs avec qui j'ai eu la chance de directement collaborer, ce travail doit beaucoup aux nombreux échanges que j'ai pu avoir au cours de ma thèse avec d'autres chercheurs, principalement au cours des (trop ?) nombreuses conférences auxquelles j'ai eu l'opportunité de présenter mes travaux. Je tiens en particulier à exprimer toute ma gratitude à ceux qui ont pris le temps de lire en détail certains de mes travaux et de m'avoir fait partager leurs commentaires

et suggestions, à savoir Daniel Hausman sur le chapitre 1, Andy Denis, John Davis et Dorian Jullien sur le chapitre 2, ainsi que André Lapidus, Adrien Lutz et Jean-Pierre Potier sur une version préliminaire du chapitre 4. Je remercie également Mauro Boianovsky, Mikaël Cozic, Jurgis Karpus, Jean-Sébastien Gharbi, Cyril Hédoin, Tom Juille, Larry Samuelson et Jean-Christophe Vergnaud pour leurs commentaires et suggestions sur différentes parties de ce travail. La plupart de ces rencontres ont été rendues possibles par mes nombreux déplacements à l'étranger, et je tiens à m'excuser auprès de l'atmosphère et des générations futures pour les plus de 24 tonnes de CO2 que ces déplacements ont générées.

J'ai pu bénéficier au cours de ma thèse d'un excellent environnement de travail au sein du département d'économie de l'École Polytechnique, ainsi qu'au sein de département d'économie de University of East Anglia où j'ai passé ma première année de thèse en 2012-2013. Je souhaite tout d'abord exprimer mes plus vifs remerciements à Ophélie Doucet, Eliane Madelaine, Chantal Poujouly, Lyza Racon et Sri Srikandan, grâce à qui j'ai pu effectuer ma recherche dans les meilleures conditions administratives et informatiques. Au delà de ces considérations matérielles, je les remercie pour leur accueil au sein du département d'économie ainsi que pour leur convivialité. Je tiens également à remercier les différents chercheurs avec qui j'ai pu discuter de ma recherche au sein de ces deux institutions, à savoir notamment Raicho Bojilov, Yukio Koriyama, Jean-François Laslier, Eduardo Perez, Jean-Pierre Ponssard et Anders Poulsen. Au delà du caractère purement académique de ces années passées à Palaiseau et Norwich, c'est probablement grâce à mes camarades doctorants que cette thèse a pu se dérouler dans les meilleures conditions, en m'offrant notamment une source d'inspiration inépuisable pour les nombreuses illustrations qui peuplent cette thèse, qu'il s'agisse des choix d'épargne de certains, des paris sportifs de d'autres, ou encore de la réservation d'un avion pour une conférence au bout du monde. Pour nos soirées Académiques, les parties de pétanque, les dégustations de chocolat, les cafés aléatoires du Bôbar, et bien plus encore, je tiens à remercier les désormais docteurs Bora Erdamar, Vanina Forget, Antonin Macé, Hypatia Nassopoulos, Ali Ihsan Ozkes et Rafael Treibich, les presque docteurs de ma génération Thomas André, Jeanne Commault, Esther Delbourg, Arnaud Goussebaille, Alda Kabré, Yang Liu et Gwenael Roudaut, et enfin tous ceux qui peuvent encore pleinement profiter des joies de la thèse, Reda Aboutajdine, Faddy Ardian, Sebastian Franco Bedoya, Margot Hovsepian, Alena Kotelnikova, Christine Lavaur, Alexis Louaas, Hugo Molina et Julie Pernaudet. J'adresse finalement un remerciement tout particulier à Róbert Somogyi, avec qui j'ai partagé un directeur de thèse en plus de quelques conférences au soleil.

Cette thèse est aussi le résultat d'un long parcours académique au sein de différentes institutions qui ont contribué à ma formation en économie. Je souhaiterais tout d'abord remercier Mostafa Settaf pour m'avoir donné goût aux questions d'économie normative et de philosophie politique dès la classe prépa. Je tiens également à remercier Nicolas Drouhin et Sabine Sépari pour leur soutien au cours de ma scolarité à l'ENS de Cachan. Enfin, mon intérêt pour les problématiques liées au choix rationnel et à la modélisation des comportements coopératifs en théorie des jeux trouve son origine dans mon travail sur l'équilibre de Berge, au cours de mon stage de master au CIRED en 2011. Je tiens donc à remercier Tarik Tazdaït ainsi que Pierre Courtois pour leur encadrement, et pour avoir indirectement influencé le présent travail de manière significative.

Je profite de ces quelques lignes pour remercier les personnes qui m'ont entouré au cours de ces dernières années, et qui ont participé d'une manière ou d'une autre à la réussite de cette thèse. Pour de nombreuses raisons qui vont au delà de ce travail de thèse, je remercie tout d'abord Alain, Catherine et Tugdual pour leur affection et soutien au cours des années passées. Je remercie également Sophie Dupaquier et Arnaud Dars, Marine Salès et Toni Juet, Wafa Ben Khaled, Antoine Piétri, Marion Sierra et Étienne Boisseau, Théophile Krosi-Douté, Chloé Ramet, mes éphémères colocataires Mickaël Terrien, Jordan Trombetta et Guillaume Denis, ainsi que mes collègues cachanais Maylis Bechetoille, Aurélie Dard, Brice Fabre, Valentina Girotto, Antoine Hémon, Lionel Lecesne et Amélie Bonnet, Guillaume Monchambert, Jie Pan.... Cette liste ne pourrait être complète si je ne remerciais pas mes camarades des Planches à Musique, pour nos spectacles, vacances, randonnées, et autres festivités passées et à venir : Mathias Szpirglas et Céline Lamade (et leur compagnie enchantée), Émilien Chapon, Tom Kristensen (et ses homonymes), Jérémy Neveu (et ses grandes oreilles), Cécile Repellin, Matthieu Lefrançois (et ses lumières), Mélanie Chevance, Rouge (aka Baptiste Morisse) et Jaune (aka Benjamin Moubêche), Pat (aka Pat), Pierre Lissy (et son squash dialectique), Rémy Soucaille, Grégoire Binois (et son abbaye vidée de son plein), Camille Cullin, Noémie Vergier, Marc-Antoine Martinod (et son ministère), Aurélien Sourie, Alexis Brenes, Pierre Géhanne, Nathanaël Héron (et sa Chartreuse), Audrey Chatain, Marina Masselot, Océane Sipan, Marjorie Zuber (et son camino trop stylé), Gabriel Beauvallet (et sa toute première symphonie), Flora Borne, Stéphanie Lebrun, Claire Millon, Louise Mousset, Ivan Moyano Garcia (et son Puérro Magico), et tous ceux que j'aurais pu oublier. Je conclurai cette liste en remerciant chaleureusement mes colocataires Florent Barret (et son bêlant troupeau), ainsi que Sylvain 'Buddy' Ravets (et ses macarons earth-shattering).

# Contents

# Introduction

The object of this thesis is to study the implications of behavioural economics for normative economics. As an illustration of the kind of situation we will consider in this work, imagine:

> **Oscar's savings choice**: when hired for his first job, Oscar — an optimistic individual — had to decide how much he wanted to save for his retirement. He was then convinced that he would quickly get a higher salary, and therefore — so as to smooth his consumption over time — preferred to consume a relatively large proportion of his income at the beginning of his career. He knew that this strategy could be risky, and therefore planned to decrease his level of consumption in the future if he could not manage to get a better-paid job. Unfortunately for him, Oscar never earned a significantly higher salary, and — due to inertia in his consumption habits — was also unable to keep his initial commitment. He then ended up with a quite low old age pension. Oscar now regrets his past decision and thinks that, were he able to turn back the clock, he would change his choice, based on better information about his ability to keep commitments and not misled any more by his optimistic expectations.

Suppose that — as a behavioural economist — you know that many individuals, like Oscar, are too optimistic concerning their beliefs about getting a higher salary, but also about keeping their commitments. Since they are likely to regret their choice once retired, you also know that they are probably doing a mistake when saving little for their retirement. What should be done in this situation? Should you let Oscar make his own choice, knowing that he is likely to regret it? Or should you interfere with Oscar's choice one way or another — in his own interest?
A common answer to this kind of problem would be to invoke J.S. Mill's *harm principle*, stating that 'the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others. His own good, either physical or moral, is not a sufficient warrant'. Although Oscar's choice gives you good reasons 'for remonstrating with him, or reasoning with him, or persuading him, or entreating him' (Mill, 1859, p.135), Mill states that is is not a sufficient reason for *compelling* him to save more. However, Oscar seems to be in a situation in which he would change his behaviour if he were not so optimistic: he would therefore implicitly agree with a paternalistic

intervention, since it seems that saving more is what he *truly* wants. Conly (2013) argues for instance that our poor instrumental reasoning prevents us for achieving our own goals: it can therefore be in our interest to accept that the government may prevent us from acting in accordance with our decisions. Conly indeed claims that Mill's defence of individual sovereignty over one's actions (when they do not cause harm to others) is grounded on a unrealistic account of human psychology: Mill 'overestimated the degree to which we would, if left to our own devices, actively and effectively pursue our own happiness', and 'underestimates the power of inertia and the resistance people have in recognizing that a particular course of action that they are engaged in actually is making them worse off' (Conly, 2013, p.9). Since we are likely to fail to reach happiness, due to our relatively poor reasoning abilities, it is in our interest to let the government — which is in a 'relatively objective position' — 'intervene in ways that help us reach our own, individual goals better than we would if left to our own devices' (p.10). The utilitarian defence of the harm principle seems therefore highly questionable, since boundedly rational individuals are likely to make very poor decisions that may go against their own interest.

The conceptual difficulty of the Oscar-case is that standard welfare economics has nothing to say about whether an intervention is justified or not. Welfare economics is indeed not well-suited to assess economic decisions that the agents would change had they 'complete information, unlimited cognitive abilities, and no lack of self-control' (Sunstein and Thaler, 2003, p.1162). There is indeed no consensus about how to provide normative assessments in presence of incoherent preferences (McQuillin and Sugden, 2012b): the core issue of the problem of how to reconcile normative and behavioural economics — labelled by McQuillin and Sugden (2012b) as the 'reconciliation problem' — is that the interpretation of the concept of 'preferences' in normative economics, when those preferences are not necessarily coherent (as behavioural economics suggests), is not self-evident.

## The reconciliation problem

### Preferences and welfare

Economic analysis is generally separated into two different branches: positive economics seeks to explain and describe how economic agents behave, while normative economics aims at evaluating economic outcomes, policies, and institutions. Both approaches traditionally attribute to economic agents coherent preferences, i.e. pref-

erences that are consistent and context-independent. Consistency is defined by the weak axiom of revealed preferences[1], and means that the revealed preferences of the individual constitute a complete and transitive relation, that determines her choice (Hausman, 2012, p.26). Context-independence means that individual preferences remain stable across different *contexts*. Suppose for instance that I prefer apples to oranges on Monday. Then I should also prefer apples to oranges on Tuesday: the day of the week is an element of the context within which my choice is made, and should not affect my preferences between apples and oranges — it is indeed an *irrelevant* feature of the choice problem[2]. Individual behaviour is then described by assuming that the agents behave as is seeking to satisfy their preferences, and economic outcomes are desirable to the extent that individual preferences are satisfied. However, although the concept of preferences is central in economics, little is said about what preferences actually are (Hausman, 2012, p.11): little is therefore said about the normative implications of the criterion of preference satisfaction. Hausman (2012, pp.1-2) for instance suggests four main interpretation of the word 'preferences' for English-speakers:

1. *Enjoyment comparisons*: saying that Anna prefers $x$ to $y$ means that Anna enjoys more $x$ than $y$. Preferences as enjoyment comparison are typically a matter of taste, such as preferring coffee to tea.

2. *Comparative evaluations*: saying that Bob prefers $x$ to $y$ means that Bob judges $x$ as better than $y$ in some regard (either according to a specific criterion, or to any relevant criterion). Bob may for instance prefer drinking his coffee without sugar because it is healthier (even if he prefers — in terms of enjoyment comparison — drinking his coffee with sugar).

3. *Favouring*: if a political party defends a policy of 'national preference' in terms of employment, then, *ceteris paribus*, a native has a higher chance of being hired than an immigrant; a specific class of individual is therefore preferred (or favoured), but without reference to an enjoyment comparison or a comparative evaluation.

4. *Choice ranking*: saying that Clara prefers $x$ to $y$ means that if she is faced with a choice between $x$ and $y$, she will choose $x$. If the waiter asks Clara whether she prefers coffee or tea, he only wants to know what her choice is —

[1] The basic idea of this axiom is that, if an agent chooses $x$ when $y$ is available, then the agent should never choose $y$ from a set of alternatives that includes $x$.

[2] We will discuss in more details the difficulties of the notion of 'context' in chapter 1.

and does not ask her to provide a ranking in terms of enjoyment comparison or comparative evaluation.

Saying that preference satisfaction matters therefore does not mean the same thing whether we are thinking of preferences in terms of enjoyment comparisons, comparative evaluations or choice ranking[3].

Suppose firstly that preferences are defined as enjoyment comparisons. From a descriptive perspective, it is assumed that I choose the action that best satisfies my preferences, i.e. I choose the action that maximises my 'happiness'. In this situation, preference satisfaction matters if and only if the *experience* of enjoyment, in line with Bentham's utilitarianism, is valuable for itself (see for instance Layard (2005)).

Suppose now that preferences are defined as comparative evaluations. My preferences therefore represent my own conception of my *well-being*: although I have a reason to prefer my coffee with sugar (because I enjoy the taste of sugar), and also a reason to prefer my coffee without sugar (because it is better for my health), my preference for one option over another depends solely on the weights $I$ give to each reason. When assuming that individuals are characterised by coherent and stable preferences, we therefore implicitly assume that they have a coherent and stable conception of their own well-being. Hausman (2012) defends this reason-based account of preferences in economics, by defining preferences as a 'total subjective comparative evaluation'. He indeed argues that, in economics, 'to say that Jill prefers $x$ to $y$ is to say that when Jill has thought about everything she takes to bear on how much she values $x$ and $y$, Jill ranks $x$ above $y$' (p.34). Preferences are therefore comparative evaluations, are expressed in terms of subjective value, and are total since they consider the totality of the factors that the individual considers to be relevant for her choice. Preferences are therefore understood as the product of reasoning, and should be considered 'more as judgments than feelings' (p.135). Preference satisfaction therefore matters if and only if achieving well-being (defined as a total subjective comparative evaluation) matters.

Finally, suppose that preferences are defined as a choice ranking. Saying that preference satisfaction matters only means here that I should be able to choose what I want to choose when I want to choose it. This interpretation differs from the two others since it does not refer to a welfarist criterion (either objective, in terms of happiness, or subjective, in terms of self-assessed well-being): what matters from this perspective is the sovereignty of the consumer over her own choices,

[3]We do not consider the interpretation in terms of 'favouring' here, since it is not related to a comparison between different alternatives — and is therefore of little interest for our discussion.

i.e. her freedom to make the choices she wants to make at the time she wants to make them. Preference satisfaction therefore matters if and only if respecting one's freedom of choice is what ultimately matters.

A crucial assumption of the two welfarist criteria is that the satisfaction of my preferences should reveal an underlying indicator of welfare (either objectively measured in a mental-state perspective, or subjectively defined as one's total comparative subjective evaluation). Therefore, preference satisfaction may provide an acceptable normative criterion if and only if the preferences revealed though my choices are consistent and context-independent. On the contrary, this claim is unnecessary for the third interpretation of preference satisfaction in terms of consumer sovereignty. What matters in this last situation is indeed that I should be able to choose what I want to choose when I want to choose it: it is therefore not necessary to impose any formal constraint on the internal coherence of my revealed preferences.
Therefore, as long as individual preferences are assumed to be coherent, we can elude the philosophical question of why preference satisfaction matters. The 'reconciliation problem' refers to the difficulties that may arise for normative economics when preferences are not coherent, as suggested by behavioural economics.

### Preference satisfaction when preferences are not coherent

Standard economic theory is built on the assumption that people act as if seeking to satisfy coherent preferences, and are instrumentally rational given their beliefs and those preferences. Behavioural economics however highlighted the existence of numerous and systematic inconsistencies in individual preferences (e.g. Kahneman and Tversky (2000) on framing effects, Frederick et al. (2002) on time inconsistencies, Camerer (2003) and Crawford et al. (2013) on strategic interactions, Sobel (2005) on interdependent preferences). Those evidence raise serious issues for normative economics, since it is not certain any more that the satisfaction of individual preferences remains a valid normative criterion.

Suppose firstly that the normative criterion is consumer sovereignty, i.e. preference satisfaction when preferences are defined as choice rankings. Whether individual preferences are coherent or not is not an issue: since what matters is one's freedom of choice, we do not need to pay attention to the effective choice of the individuals, but only to the opportunity they had making choices among a wide

range of alternatives. This argument has been developed by Sugden (2004, 2007), who suggests using a normative criterion of opportunity rather than preference satisfaction: what matters is the ability for the individual of having at her disposal a wide range of available actions, and not the consequences of the choice she eventually makes.

Suppose now that the normative criterion is happiness (or more generally any hedonistic experience that can be objectively measured[4]), as endorsed for instance by Kahneman et al. (1997). Since the preferences I reveal through my choices can be incoherent, they do not provide a good indicator of my level of happiness. Satisfying my preferences is therefore not necessarily desirable in itself: it is then possible to argue that behavioural findings justify paternalism. McQuillin and Sugden (2012b, p.558) however note that this conclusion is not self-evident: following Berg and Gigerenzer (2010), it is also possible to interpret deviations from rational choice theory axioms as the result of the application of successful heuristics (i.e. heuristics adapted to the environment within which the individuals interact) that differ from payoff maximisation. It is therefore not obvious that people would benefit from conforming to the neoclassical model of behaviour (this point will be extensively discussed in chapter 7 — we will indeed characterise the set of games for which it is 'ecologically' rational to deviate from payoff maximisation).

The most common response given by behavioural economists to the reconciliation problem however rests on the interpretation of preferences as comparative evaluations. This approach was introduced by Sunstein and Thaler (2003) and Camerer et al. (2003), and later popularised by Thaler and Sunstein (2008) thanks to their book *Nudge: Improving Decisions about Health, Wealth, and Happiness*. The proponents of this criterion defend a 'soft paternalism' – 'libertarian paternalism' for Sunstein and Thaler and 'asymmetric paternalism' for Camerer et al. (2003) – and justify paternalistic interventions if they can help the individuals to get what they would have chosen on their own if they were rational. Sunstein and Thaler defend this paternalistic position by arguing that paternalism is actually 'inevitable' (Sunstein and Thaler, 2003, p.1171). People are highly sensitive to framing effects, and their decisions are influenced by features of the environment of choice that are not relevant from the perspective of the social planner (Bernheim and Rangel, 2009, p.55), or in other words, by features they would not consider to be relevant if they

[4]Although providing an objective measure of experienced happiness remains a delicate task, important progresses have been made in developing operational measures of happiness. See Frey and Stutzer (2002) for a review.

were rational. Since there is no 'neutral' frame, any way of presenting a situation of choice will necessarily make an option more salient than another one (by defining for instance a default option, a first option in a list of alternatives, etc.). Sunstein and Thaler argue then that the individual in charge of the design of the situation of choice – the *choice architect* – should choose the choice architecture so as to help people to improve their well-being, 'as judged by themselves' (2008, p.5). The logic of libertarian paternalism (henceforth 'LP') is therefore that people aim to choose the options that will make them better off, but that – due to human fallibility – they often make non rational choices, and miss their objectives. Since a choice architect has the possibility of slightly influencing people's choices, she should *nudge* them so that they achieve *in fine* their goals. A 'nudge' is then defined as 'any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any option or significantly changing their economic incentives' (Thaler and Sunstein, 2008, p.6)[5]. Since the individuals are not forced to choose the option the choice architect wants them to choose, Sunstein and Thaler argue that nudges preserve individuals' freedom of choice. We will borrow Bernheim and Rangel (2009) notion of *behavioural welfare economics* (henceforth 'BWE') to designate the body of literature that intends to produce welfare assessments under the claim that individuals may make suboptimal decisions in terms of their self-assessed well-being.

The two central claims of LP are therefore that (i) we can make the individuals better off, 'as judged by themselves', since they can suffer from making mistakes, and (ii) we can do so while preserving their freedom of choice. The first claim characterises the paternalistic dimension of nudges, while the second claim ensures that nudges are compatible with libertarian principles. Several authors however question the claim that nudges are actually paternalistic, such as Hausman and Welch (2010, p.136), who argue that most of the nudges defended by LP are cases of rational persuasion. Since the choice architect is not supposed to impose her own normative views to the individual, and that individuals are not constrained by the choice architect, they claim that Sunstein and Thaler definition of 'paternalism' is unsatisfactory — LP would be better understood as *beneficence*. Thaler and Sunstein (2003, p.175) indeed define a policy as paternalistic if 'it is selected with the goal of influencing the choices of affected parties in a way that will make those parties better off': LP is therefore a *means* paternalism rather than an *ends* paternalism (Sunstein, 2014b, pp.19-20). Unlike ends paternalists who pursue objectives that are

[5]Note however that Sunstein and Thaler provide a slightly different definition a few pages later, as 'any factor that significantly alters the behavior of Humans although it would be ignored by Econs' (p.8). See Mongin and Cozic (2014) on the different interpretations of 'nudges'.

different from the ends of the individuals, means paternalists want to help people to achieve their own ends. Furthermore, LP is a form of *soft* rather than *hard* paternalism, since it does not impose to the individuals the choice to make so as to achieve their own ends. Since nudges do not limit the freedom of the individuals — and that limitation of freedom constitutes one of the main reasons why paternalism is morally problematic (Feinberg, 1971, Dworkin, 1972) — nudges are seen as the ideal policy to deal with boundedly rational individuals:

> We might even venture a more general principle, which might be called the First (and only) Law of Behaviorally Informed Regulation: *In the face of behavioral market failures, nudges are usually the best response, at least when there is no harm to others.* (Sunstein, 2014b, p.17, emphasis in original)

If we consider the Oscar-case, then LP recommends nudging Oscar today such that he will benefit from higher savings when retired. This can be done by exploiting his bounded rationality through framing effects: we could for instance design a default option such that he will be inclined to save more without significantly limiting his freedom of choice (Choi et al., 2004, Thaler and Benartzi, 2004). Note that the main difference between LP and BWE is that, although both approaches promote means paternalism, only LP explicitly recommends a soft paternalism (more traditional interventions such as taxing unhealthy food are for instance justified by BWE, although they would not be supported by LP).

## Outline of the thesis

The object of this thesis is twofold. Our first objective is to offer a methodological and philosophical discussion of behavioural welfare economics, so as to determine whether it provides a satisfying solution to the reconciliation problem. We show that, so as to be coherent, the proponents of BWE must accept a problematic model of agency, according to which an individual can be conceptualised as an inner rational agent trapped in an outer psychological shell. This model is grounded on the hypothesis that individuals are defined by latent coherent preferences: this hypothesis is however not properly justified and raises serious methodological and philosophical difficulties. We argue instead that the normative challenge raised by behavioural economics is that individuals may lack of autonomy: rather than considering that what matters is the satisfaction of one's 'true' preferences, we argue that what matters is the ability to choose one's own preferences.

We argue that BWE does not offer an adequate solution to the reconciliation problem, because its underlying model of preferences — which derives from Pareto's analysis of logical actions — distinguishes between *true* (and therefore normatively valuable) preferences and *mistakes*, without precisely clarifying the status of those 'mistakes'. The second object of this thesis is therefore to develop an alternative model of preferences compatible with behavioural findings that does not rely on a primitive in terms of true, subjective preferences. Our model is grounded on Bacharach's variable frame theory and explicitly considers the possibility for the agents to form to some extent their own preferences. Preference incoherences result from the existence of different frames through which the individual can represent the choice problem, but are not defined as deviations from a single coherent preference ordering. In addition of providing a model of endogenous preferences that does not rely on a notion of true preferences, we also consider the possibility of collective agency, i.e. that agents can conceive themselves and act as the members of a group — they are therefore able to *team reason*.

## Part I: A Rational Self Trapped in a Psychological Shell

The first part of our thesis discusses the accuracy of behavioural welfare economics as a solution to the reconciliation problem. Our main argument is that BWE is a viable solution if and only if we accept that people are fundamentally defined by true preferences, i.e. preferences that would determine their choices, if they were rational. The central argument of BWE is indeed that people fail to choose what they *truly* want, and therefore that they should be helped so as to achieve their own ends. The difficulty of this approach is that little is said about what 'to truly want' means: we argue in particular that it is not possible to define unambiguously what someone truly wants (chapter 1). We suggest that this idea of true preferences can be traced back to Pareto's work and his definition of the *Homo economicus* (chapter 2): BWE can then be understood as Pareto's project of 'reconstructing' the individual for applied economics, once it has been decomposed into different *Homines* for the study of pure economics (Pareto, 1909, chap. 1, §26). BWE then defines *Homo economicus*'s preferences as one's true preferences: it is therefore assumed that what I truly want is what the *Homo economicus* would do (chapter 3). We finally suggest that the reason why economists intend to give to people what they truly want is the result of the third-person perspective they endorse when providing normative assessments. This perspective may however offer a biased diagnosis of the normative issues faced by boundedly rational individuals (chapter 4).

Chapter 1 reviews the different works belonging to the programme of behavioural welfare economics. Those works treat deviations from conventional rational choice theory as *mistakes*, and assume that we can 'purify' revealed preferences so as to discover the underlying true preferences of the individual. This *preference purification* approach implicitly uses a dualistic model of the human being, in which an inner rational agent is trapped in an outer psychological shell. This model is philosophically and psychologically problematic, since it requires the existence of a latent mode of reasoning that can generate a unique complete and context-independent preference ordering. We argue in particular that nothing rationally requires preferences to be context-independent: considering that context-dependent preferences are the result of errors of reasoning is fundamentally misleading.

Chapter 2 then investigates the historical origins of the notion of true preferences. We suggest that the marginalist revolution generated two distinct trends in economics: the first one (with Jevons, Menger, Edgeworth, Pantaleoni and later the Austrian school) was interested in the study of human actions — as a means to explain the phenomenon of exchange in society — while the second (with Walras, Marshall and Pareto) was directly studying the institution of markets, i.e. intended to explain *how* rather than *why* people do exchange. Each tradition developed a distinct conception of the economic man, as the idealisation of an economic agent in the former, and as a representative individual in the latter. We argue that Pareto intended to study a representative agent, and therefore was able to get rid of psychological considerations in his definition of the economic man. The ambiguity of his definition of the *Homo economicus* however led many economists to consider it as a model of individual behaviour, and therefore led them to assume that individual behaviour may be approximated by the behaviour of a rational *Homo economicus*: the individual was then conceived as a rational agent who progressively frees herself from her psychological shell, and who progressively tends to satisfy her true preferences. We however suggest that the assumption of coherent preferences postulated for studying economic agents in repeated markets was probably a property of the institution itself of repeated markets, rather than of individual preferences.

Chapter 3 then argues that the development of behavioural welfare economics can be seen in the direct continuation of Pareto's work. Pareto indeed suggested

that, while pure economics is only interested in the behaviour of the *Homo economicus*, applied economics requires to model individuals as they really are, since the model of the *Homo economicus* does not generally offer a relevant description of individual behaviour. So as to provide sound policy recommendations, welfare economists should therefore 'reconstruct' the individual from the different *Homines* we separated for theoretical study. Behavioural welfare economics may then be interpreted as the attempt to reconstruct the individual, as the aggregation of a rational agent, the *Homo economicus* with coherent preferences, and a 'psychological' agent, the *Homo psychologicus*, whose preferences are generally incoherent. Although Pareto would probably have agreed so far, BWE goes a step further by assigning to the satisfaction of the *Homo economicus*'s preferences a normative value. In addition of being able to isolate the true preferences of the individual (the preferences of her *Homo economicus*), it is assumed that the satisfaction of those specific preferences matters. BWE therefore claims that (i) we can associate to each agent a unique underlying consistent and context-independent preference relation, and that (ii) the satisfaction of those true preferences is the normative criterion.

Chapter 4 concludes this first part by questioning the second claim of BWE, i.e. that the objective of public policies should be the satisfaction of one's purified preferences. This claim indeed justifies the paternalistic dimension of libertarian paternalism, since the individuals are not able to satisfy their true preferences by themselves. We argue that the argument according to which the satisfaction of one's purified preferences matters implicitly assumes that individuals are just passive *loci* of experience: BWE denies the status of *agents* for individuals, and in particular that individuals can actively contribute to the shaping of their own preferences. This is related to the third-person perspective economists endorse when providing normative assessments: they indeed take the standpoint of an omniscient and omnipotent social planner, whose objective is to design the society such that the individuals achieve *in fine* their own ends (or, within the context of BWE, the position of a choice architect who designs the choice architecture such that the individuals satisfy *in fine* their true preferences). On the contrary, accepting that individuals are more than mere preference relations implies that what matters is not necessarily preference satisfaction, but the agents themselves. We therefore defend a normative criterion in terms of individual autonomy, according to which it is the ability to choose and accept one's preferences that matters.

### Part II: A Model of Endogenous Preferences

We show in part I that the main difficulties of BWE are related to its implicit commitment to a problematic model of preferences, according to which individuals are defined by true preferences whose satisfaction is normatively desirable. This model is indeed philosophically and psychologically questionable (chapter 1, 3), an investigation into its historical origins show that it can only be used to study competitive markets (chapter 2), and it may offer a biased perception of the normative issues raised by behavioural findings (chapter 4). The object of this second part is therefore to provide an alternative model of preferences for normative analysis, that may integrate findings from behavioural economics. Our model relies on Bacharach's variable frame theory and on the theory of team reasoning. In a nutshell, we assume that the different states of the world resulting from individual choices cannot be unambiguously described, i.e. there are several equally valid perceptions of the same state of the world. In particular, there does not exist any perception with a higher normative value than another (as it would be implied with BWE's notion of true preferences). According to their perception of the game (their *frame*), the individuals may decide to switch to another perception as a means to satisfy what they perceive as being their interest in their initial frame. We show in particular that individuals are likely to adopt what Bacharach calls a 'we-frame', i.e. to consider themselves as the member of a group and to be actuated by the group objective. Players strategising with a we-frame are then *team reasoning*.

Chapter 5 firstly questions an implicit but never justified claim of BWE, i.e. that individuals would be better off if they were behaving as in standard rational choice theory. Although this seems to be a reasonable assumption when studying individual decisions, this claim is generally false within the context of strategic interactions. We develop a model of endogenous preferences, that will constitute the basis of the general model of preferences we develop in chapter 6. We assume that players are able to choose in a precommitment game the weights they assign to the other players' material payoff in their own preferences and determine the optimal weights each player should choose so as to maximise *in fine* her own material payoff. We highlight a systematic relation between the supermodularity (submodularity) of the game and the formation of cooperative (competitive) preferences. This suggests that the strategic nature of the institution within which individuals interact may strongly influence their own preferences: as an application, we investigate the possibility to shape individual preferences thanks to public policies, by altering the strategic nature of the game. We show that, in the case

of climate change negotiations, international agreements relying on technology standards with trade sanctions rather than objectives of pollution abatement are more likely to succeed. Such agreements indeed create a coordination game and cut the strategic substituability of the initial game — that would have given incentives to adopt more competitive preferences.

Chapter 6 develops a formal framework of team reasoning, based on Bacharach (1999)'s work: we reframe Bacharach's model as a generalisation of Bayesian games, and develop a strategic model of collective preferences. We indeed argue that individuals engaged in team reasoning are actuated by the collective intention of satisfying their individual preferences, and we show that this procedure can be represented as the satisfaction of a specific class of collective preferences. We highlight that, in the presence of several groups, team reasoners are likely to develop aggressive preferences towards outsiders in games presenting strategic substitutes, while strategic complementarities will lead to cooperation between groups. We finally show that, under the assumption of common knowledge of rationality, rational players should be able to form commitments and to choose to team reason: for almost every game, it is actually in the interest of a rational player to identify with a group, so as to be able to choose collective preferences that may *in fine* benefit her as an individual.

Chapter 7, in line with Bacharach's project of justifying in an evolutionary perspective the emergence of team reasoning, studies the ecological rationality of team reasoning. We firstly argue that evolutionary game theory should study the evolution of individual heuristics rather than of individual strategies or preferences, since it may provide more complete and insightful predictions. We define the $\phi$-core as a refinement of the $\gamma$-core by assuming that the deviating coalition acts as a Stackelberg leader and the remaining individuals as singletons. We show that maximising one's payoff is evolutionary stable if and only if the resulting Nash equilibrium belongs to the $\phi$-core. We can then highlight that, when the players use the heuristics that outperform payoff maximisers, they behave as if they were team reasoning. This result suggests that, from an evolutionary perspective, team reasoning is ecologically rational: although individual behaviour differs from the prescription of the 'constructivist rationality' of neoclassical economics, individual heuristics consistent with team reasoning are well adapted to the environment within which the players interact.

# Part I

# A Rational Self Trapped in a Psychological Shell

Chapter 1

# Preference Purification and the Inner Rational Agent*

Contents




**Abstract:** Neoclassical economics assumes that individuals have stable and context-independent preferences, and uses preference-satisfaction as a normative criterion. By calling this assumption into question, behavioural findings cause fundamental problems for normative economics. A common response to these problems is to treat deviations from conventional rational-choice theory as mistakes, and to try to reconstruct the preferences that individuals would have acted on, had they reasoned correctly. We argue that this preference purification approach implicitly uses a dualistic model of the human being, in which an inner rational agent is trapped in an outer psychological shell. This model is psychologically and philosophically problematic.


*This paper is a joint work with Gerardo Infante and Robert Sugden, forthcoming in the *Journal of Economic Methodology*.

## 1.1 Introduction

In neoclassical economics, it is standard practice to assume that individuals have stable and context-independent preferences over all economically relevant outcomes, and to use the satisfaction of those preferences as the primary normative criterion. By calling this assumption into question, the findings of behavioural economics are causing fundamental problems for normative economics. In this chapter, we critically evaluate a response to these problems that has been advocated by many prominent behavioural economists, and that has recently been endorsed by Hausman (2012) in a philosophical enquiry into how the concepts of preference, value, choice and welfare are (and ought to be) used in economics. Following Hausman (p. 102), we will call this approach 'preference purification'. The essential idea is that when an individual's decisions are inconsistent with defensible assumptions about rational choice, those decisions can be treated as mistakes. The task for welfare economics is then to reconstruct the preferences that the individual would have acted on, had her reasoning not been distorted by whatever psychological mechanisms were responsible for the mistakes, and to use the satisfaction of these reconstructed preferences as a normative criterion.
We will argue that this approach implicitly uses a dualistic model of the human being, in which an inner rational agent is trapped inside a psychological shell. The inner agent is pictured as the locus of the identity of the human being and as the source of normative authority about its interests and goals. There is no attempt to represent the psychology of this agent; its rationality is simply taken as given. The psychological mechanisms that induce deviations from supposedly rational choice are treated as properties of the outer shell that can prevent the inner agent from achieving its objectives. Whether viewed in the perspective of psychology or of philosophy, this model is problematic.

We will begin by describing some of the context-dependent features of real decision-making behaviour that cause problems for conventional welfare economics (section 1.2). We will explain the preference purification approach that behavioural welfare economists have used to try to resolve these problems (section 1.3), and consider Hausman's endorsement of this approach (section 1.4). Drawing on Hausman and Welch (2010)'s attempt to define a concept of autonomy that is appropriate for behavioural agents, we will explain the sense in which the preference purification approach presupposes the model of the inner rational agent (section 1.5). We will argue that the idea that context-dependent choices are caused by errors of reasoning is fundamentally misconceived (section 1.6). Finally, we will offer a conjecture about

why behavioural economists have been attracted by the model of the inner rational agent (section 1.7).

## 1.2 Background

Our main focus will be on a class of cases that feature prominently in discussions about the normative significance of behavioural findings. These are cases in which a person's preferences, choices or judgements are strongly affected by factors that work through well-understood psychological mechanisms but seem to have little or no relevance to that person's well-being, interests or goals. Although there is a clear sense in which the choices made (or preferences revealed, or judgements expressed) by the person in different contexts are inconsistent *with one another*, it is not at all obvious *which* (if any) of these choices is correct — or even how 'correctness' should be defined.

Here is a typical example. In an experiment reported by Kahneman et al. (1990, pp.1338-1339), student subjects reported their valuations for coffee mugs. Subjects were randomly assigned to experimental treatments. In one treatment, each subject was asked to consider each of a range of amounts of money, and to say whether she would choose to have a mug or the money. In another treatment, each subject was first given the mug, free of charge, and then asked whether she would choose to sell it back to the experimenters at each of a range of prices (the same range of money amounts as in the first treatment). Notice that, defined in terms of what a subject can take away from the experiment, the problems faced by the two sets of subjects are exactly the same: the only difference is whether the problems are framed as *choosing* between the mug and money, or as *selling* the mug. However, the median valuation of the mug in the selling treatment ($7.12) was more than double that in the choosing treatment ($3.12). This effect can be explained by the hypothesis that losses have greater psychological salience than equal and opposite gains. (In the choosing treatment, subjects are thinking about *gaining* the mug, while in the selling treatment, they are thinking about *losing* it.) It would be very difficult to argue that the difference between being told that you have been given a coffee mug and being told that you can choose to be given one is a good reason for a two-fold difference in your valuation of the mug, and in this sense the effect seems irrational; but that does not answer the question of whether $7.12 is an irrationally high valuation or whether $3.12 is an irrationally low one.

Here is another example. Read and van Leeuwen (1998) report a field experiment in which workers made choices between free snacks which would be delivered at

a designated time a week later. The menu from which subjects could choose contained healthy options (e.g. apples) and unhealthy ones (e.g. Mars bars). There were four treatments, defined by two different times of day — 'after lunch time' and 'in the late afternoon' — at which the choice was made and (independently) at which the snack would be delivered. The background assumption was that most workers would be hungrier at the later time. Read and van Leeuwen found that, holding constant the time of delivery, subjects were more likely to choose unhealthy snacks if they made the choice in the late afternoon. In broad terms, the psychological mechanism behind this result is easy to understand. The hungrier you feel, the more attention you give to cues that are directed towards the satisfaction of hunger, and the more vividly you can imagine experiencing feelings of hunger in other situations. Thus, the hunger-satisfying properties of the Mars bar are perceived more vividly in the late afternoon, irrespective of when it will actually be eaten. Given the familiarity of the snack options and the predictability of daily fluctuations in hunger and satiation, it would be implausible to claim that differences in the time of day at which the decision is made provide good reasons for different choices about what to eat at a given time seven days later. In this sense, the context-dependent preferences revealed in the experiment seem irrational. But that does not answer the question of whether, in any given situation, it is more rational to choose an apple or a Mars bar.

Our third example concerns a less obvious principle of consistency. It is the version of the Allais Paradox discussed by Savage (1954, pp.101-103). Respondents are asked to imagine two different situations, in each of which there is a choice between two gambles. In Situation 1, the choice is between Gamble 1, which gives $500,000 with probability 1, and Gamble 2, which gives $2,500,000 with probability 0.1, $500,000 with probability 0.89, and nothing with probability 0.01. In Situation 2, the choice is between Gamble 3, which gives $500,000 with probability 0.11 and nothing with probability 0.89, and Gamble 4, which gives $2,500,000 with probability 0.1 and nothing with probability 0.9. Many people report strict preferences for Gamble 1 in Situation 1 and for Gamble 4 in Situation 2. According to the axioms of expected utility theory, a person with consistent preferences would *either* prefer Gambles 1 and 3 *or* prefer Gambles 2 and 4. But the theory does not say which of those two patterns of preference is more rational.

In this chapter we will focus on cases, like those we have just discussed, in which *choices* or *preferences* are allegedly inconsistent. However, it is important to keep in mind that *judgements* can also be systematically context-dependent in ways which do not seem to be supported by defensible reasons, and that the question of

which of the mutually inconsistent judgements is correct can be no easier to answer than analogous questions about choices or preferences. For example, people's judgements about their own happiness are subject to 'focusing illusions' that seem to result from mechanisms similar to those involved in the choice between future snacks in Read and van Leeuwen's experiment. When people are trying to judge their overall satisfaction with their lives, the implicit weights they give to different aspects of life can depend on what is currently the focus of their attention (Schkade and Kahneman, 1998) — an effect summed up in Kahneman (2011, p.402)'s maxim: 'Nothing in life is as important as you think it is when you are thinking about it'. Since most of the happiness data that economists and psychologists use are generated from self-reported judgements, one should not assume that the problem of context-dependent preferences can be resolved simply by defining 'true preferences' in terms of happiness.

We recognise that there are some cases of context-dependent choice for which the definition of a person's true preferences or best interests is fairly uncontroversial. For example, in some retail energy markets, competing suppliers offer exactly the same product, priced according to different tariffs. It seems unexceptionable to assume that, for any given quantity bought, consumers have an underlying preference for paying less rather than more. In fact, when tariffs are complex, consumers often fail to buy from the supplier offering the lowest final price (Wilson and Waddams Price, 2010). Representing such choices as mistakes, defined relative to 'true' preferences for low prices, may be a reasonable modelling strategy. In this case, however, the assumption that is taken to be uncontroversial in defining mistakes equates the consumer's *subjective* ranking of options (alternative tariffs) with an objective ranking (in inverse order of their prices) that is independent of the consumer's perceptions or judgements. There is no obvious analogue to this objective ranking in cases such as the choice between buying or not buying a coffee mug, or between eating an apple and eating a Mars bar.

The distinction between objective and subjective rankings is important because of the way in which the idea of mistakes is used in behavioural welfare economics. A recurring theme in this literature is that the findings of behavioural economics justify policies which 'nudge' individuals towards those choices that are in their best interests (e.g. Camerer et al. (2003); Sunstein and Thaler (2003) [henceforth 'ST']; Thaler and Sunstein (2008) [henceforth 'TS']). The element of paternalism in these proposals can be made more palatable by suggesting, not only that their aim is to increase the welfare of the targeted individuals, but also that welfare is being measured *according to those individuals' own judgements*, and that the choices that individuals are being nudged away from would be *mistakes*. These

suggestions are often expressed through the idea that nudges help individuals to make what, on reflection, they themselves would recognise as better choices. For example, asking the reader to consider a problem of choosing between a large number of prescription drug plans, Thaler and Sunstein (2008, p.10) say: '[Y]ou might benefit from a little help. So long as people are not choosing perfectly, some changes in the choice architecture could make their lives go better (as judged by their own preferences...)'[1]. Contrast this notion of helping people to avoid mistakes with the more overt paternalism of a parent who limits a two-year-old child's consumption of chocolate in the interests of a balanced diet. The parent believes her action promotes the child's welfare, but the child's wish to eat chocolate is not a mistake. As a real desire for an experience that really is pleasurable, it makes good sense in terms of all the reasons that the child is capable of understanding. If context-dependent choices can be represented as mistakes, the relationship between a paternalistic policy-maker and a targeted individual looks less like that between a benevolent guardian and an incompetent ward.

To avoid misunderstanding, however, we must make clear that this chapter is not primarily concerned with whether (or how far) public policy should be paternalistic. It is possible to investigate questions about individual welfare without presupposing that governments ought to adopt whatever policies can be shown to maximise well-being. Although many advocates of preference purification present it as a technique that can be used in designing paternalistic policies, this linkage is not universal. In particular, Hausman (2012) endorses preference purification as a tool for the measurement of welfare in applied economics, but has serious reservations about the use of nudges as a policy tool. By bracketing out the question of what governments ought to do with welfare measurements, we are able to evaluate Hausman's philosophical arguments for preference purification.

We also bracket out questions about the use of nudges for non-paternalistic purposes. A significant part of the nudge literature is directed at using behavioural insights to induce 'behaviour change' in situations in which the targeted individuals do not seem to be making mistakes in satisfying their own preferences or in promoting their own welfare: they are simply frustrating the achievement of some public policy objective. For example, TS's catalogue of emulation-worthy policies includes nudges designed to reduce littering, to increase registration in organ donation programmes, and (through naming and shaming polluting firms) to reduce the release of potentially hazardous chemicals into the environment (pp. 60, 175–182, 190–191).

[1] TS repeatedly describe nudges as 'helping' the individuals at whom they are directed. Looking only at their first chapter, one finds this use of 'helping' on pp. 6, 7, 9, 10, 11 and 14.

The UK Behavioural Insights Team prides itself on having designed nudges which make people more likely to pay their tax bills on time, to the benefit of the public finances (Halpern and Nesterak, 2014). Discussion of the legitimacy of such policies has focused on issues of transparency and democratic accountability (Hansen and Jespersen, 2013, Sunstein, 2014a). In contrast, our concern in this chapter is with the definition and measurement of welfare[2].

## 1.3 Behavioural welfare economics and preference purification

Since Sunstein and Thaler have been particularly influential in the development of behavioural welfare economics, we begin by looking at the role of preference purification in their arguments. Their original paper (ST) sets out a manifesto for *libertarian paternalism*[3]. Their later book *Nudge* (TS) extends and popularises the ideas in ST.

One of Sunstein and Thaler's key claims is that the findings of behavioural economics make paternalism unavoidable: the anti-paternalist position is 'incoherent'. a 'nonstarter'. In both works, this claim is developed in relation to a now-familiar cafeteria example. The premise is that customers' choices between alternative food items are influenced by the prominence with which those items are displayed on the cafeteria counter. Knowing that some items are healthier than others, the cafeteria director has to choose the relative prominence with which different items are displayed. ST consider two apparently reasonable strategies that the director might adopt: she could 'make the choices that she thinks would make the customers best off, all things considered' or she could 'give consumers what she thinks they would choose on their own'. We are told that the second option is 'what anti-paternalists would favor'. but that the anti-paternalist argument for this option is incoherent. By assumption, the customers

> lack well-formed preferences, in the sense of preferences that are firmly held and preexist the director's own choices about how to order the

[2] We will discuss in chapter 4 some ethical concerns raised by the use of nudges as a policy tool.

[3] Contemporaneously with ST, Camerer et al. (2003) advocated 'asymmetric paternalism' as a normative response to the findings of behavioural economics. There are close similarities between the two proposals. Asymmetric paternalism is presented as a way of helping boundedly rational individuals to avoid 'decision-making errors' that 'lead people not to behave in their own best interests' (pp. 1211–1212). However, Camerer et al. give even less guidance than ST about how individuals' interests are defined or how they can be identified.

> relevant items [along the counter]. If the arrangement of the alternatives has a significant effect on the selections of the customers make, then their true 'preferences' do not formally exist[4].

Sunstein and Thaler conclude that the first strategy, despite being paternalistic, is the only reasonable option for a well-intentioned director (ST: 1164–1165, 1182; see also TS: 1–3).

Notice that, as in the general class of problems described in section 1.2, the choices of the cafeteria customers are context-dependent in a way that has a psychological explanation (more prominently-displayed items are more likely to engage attention) but does not seem relevant to customers' interests or goals. Such cases are central to Sunstein and Thaler's argumentative strategy. The key innovation of libertarian paternalism is the idea that individuals' choices from given sets of (objectively defined) options can be influenced by interventions which affect only the (subjectively perceived) framing of the decision problem. Such nudges can work only in cases in which choices are context-dependent.

Notice also that the cafeteria problem is presented as a problem *for the cafeteria director*. The director is understood as someone who acts on her own authority and responsibility, but with the objective of benefiting her customers. Sunstein and Thaler describe this role as that of a 'planner' (in ST) or 'choice architect' (in TS). The idea that normative recommendations are addressed to a benevolent planner is a common device in welfare economics, and leads naturally to the further idea that those recommendations should be directed at increasing the well-being of the individuals for whom the planner is planning. In TS, this idea is given a slightly different twist: Sunstein and Thaler say that their recommendations are designed to 'make choosers better off, *as judged by themselves*' (TS: 5; italics in original). The italicised clause recurs with minor variations throughout TS (e.g. pp. 10, 12, 80). The implication, we take it, is that although the planner acts on her own responsibility, she tries to respect each individual's subjective judgements about what makes him better off.

Sunstein and Thaler's approach to normative economics requires that the planner can reconstruct each individual's judgements about his own well-being, even though

[4]We take it that, in this passage, ST are using 'preference' in the sense that it is used in conventional economic theory — that is, as a binary relation over potential objects of choice that is consistently revealed in an individual's decisions. In this sense, the cafeteria customer does not have well-defined ('true') preferences over food items. In section 1.2 above, we followed a common practice in behavioural economics by using the concept of 'true preference' is a different sense — to refer to the preferences on which (it is supposed) an individual would act in the absence of errors.

these judgements are not always revealed in his choices. But how, at the conceptual level, are we to understand these judgements? And how is the planner to reconstruct them? The closest that Sunstein and Thaler come to addressing these questions systematically is in their discussion of decision-making errors.
Immediately after presenting the principle of trying to make choosers 'better off, *as judged by themselves*', TS undertake to show that

> in many cases, individuals make pretty bad decisions — decisions that they would not have made if they had paid full attention and possessed complete information, unlimited cognitive abilities, and complete self-control. (TS: 5)

The corresponding passage in ST uses almost the same characterisation of decisions that would not have been made if individuals had been fully rational, and refers to these as 'inferior decisions in terms of their [i.e. the individuals'] own welfare' (ST: 1162). The implication is that, for Sunstein and Thaler, the criterion of individual well-being is given by the preferences that the relevant individual would have revealed, had his decision-making not been affected by *reasoning imperfections* — that is, limitations of attention, information, cognitive ability or self-control. So the task for the planner is to try to reconstruct individuals' underlying or latent preferences by simulating what they would have chosen, had they not been subject to reasoning imperfections. This is *preference purification.*

Notice that preference purification cannot provide the welfare criterion that Sunstein and Thaler need unless latent preferences are context-independent. The context-dependence of revealed preferences, with the supposed implication that paternalism is unavoidable, provides the starting point for Sunstein and Thaler's argument for libertarian paternalism. But if the choice architect's decision criterion turned out to be context-dependent too, that argument would be fatally undermined. The assumption that latent preferences are context-independent is implicit in Sunstein and Thaler's arguments, but is never defended. One of their favourite rhetorical strategies is to characterise their opponents as maintaining that human beings are not subject to reasoning imperfections — that human beings can 'think like Albert Einstein, store as much memory as IBM's Big Blue, and exercise the willpower of Mahatma Gandhi'. The reader is invited to agree with the hardly controversial proposition that 'the folks we know are not like that'. and encouraged to infer that *this must be why* ordinary folks' choices have the apparently irrational patterns that behavioural economics and psychologists have discovered (TS: 6–8).

But this inference is not as obviously valid as it may seem. We will return to this issue later, but to lay a foundation for later arguments we invite the reader to think about the following situation:

> **Joe and SuperReasoner**. Imagine a being — let us call him SuperReasoner — who has the intelligence of Einstein, the memory of Big Blue, and the self-control of Gandhi. Imagine in addition (since this is also part of Sunstein and Thaler's characterisation of perfect reasoning) that SuperReasoner's capacious memory contains every item of information that can be extracted from any existing publication or database. Otherwise, however, SuperReasoner is just like some ordinary human, whom we will call Joe. If Joe were in Sunstein and Thaler's cafeteria, his choices between food items would be influenced by the prominence of their displays. Now imagine taking SuperReasoner into the cafeteria. Would the probability of his choosing cream cake be independent of the position of cream cake on the counter?

Bernheim and Rangel (2007, 2009) propose an approach to behavioural welfare economics that is similar to preference purification. Presuming it to be self-evident that welfare economics is addressed to 'the planner'. they characterise 'standard welfare economics' as 'instruct[ing] the planner to respect the choices an individual would make for himself' (2007: 464). Their objective is to extend this form of welfare economics to cases in which choices are context-dependent. The key concept in their theoretical framework is a *generalised choice situation* (GCS) for a given individual, consisting of a set of 'objects' from which the individual must choose one, and a set of 'ancillary conditions'. Ancillary conditions are properties of the choice environment that may affect behaviour but which the planner treats as normatively irrelevant. (Applying this conceptual scheme to the cafeteria, food items are objects while ways of displaying them are ancillary conditions.) The individual's choice behaviour is represented by a correspondence which, for each GCS, picks out the subset of objects that the individual is willing to choose. Bernheim and Rangel's first line of approach is to propose a criterion that respects the individual's revealed preferences over pairs of objects if those preferences are not affected by changes in ancillary conditions, and instructs the planner 'to live with whatever ambiguity remains' (2009: 53). They then suggest that this rather unhelpful criterion might be given more bite by the deletion of 'suspect' GCSs. A GCS is deemed to be suspect if its ancillary conditions induce impairments in the individual's ability to attend to or process information, or to implement desired courses of action. In effect, this approach purifies choice data by eliminating any choices that were made when the individual's reasoning was impaired. Considering

only the purified data, it then uses the satisfaction of context-independent revealed preferences as the normative criterion. Notice that this approach yields welfare rankings only for those pairs of objects for which revealed preferences, after purification, are context-independent.

A different way of using the idea of purification is to begin by assuming the existence of context-independent latent preferences, and to propose some specific model of the psychological processes that intervene between those preferences and actual choices. Given such a model, one can then investigate how far and under what assumptions latent preferences can be reconstructed from observations of choices. Salant and Rubinstein (2008) follow this approach within a general theoretical framework similar to Bernheim and Rangel's. They define an *extended choice problem* for an individual as a pair $(A, f)$ where $A$ is a set of mutually exclusive and exhaustive alternative objects of choice, and $f$ is a 'frame'. The individual's decisions are determined by the interaction of her frame-independent 'underlying preferences' with decision-making heuristics that are activated by, and conditional on, the frame (p. 1288). Like Sunstein and Thaler, they imagine a 'social planner' who chooses the frame with the aim that the individual's choice should be consistent with her underlying preferences (p. 1294).

This approach is often used to derive normative implications from behavioural models. The model of 'salience and consumer choice' presented by Bordalo et al. (2013) is a good example. The psychological intuition behind this model is that choice options ('goods') can be described as bundles of characteristics, and that when a consumer evaluates a good, she gives most attention, other things being equal, to those attributes of that good that are perceived as most 'salient' (that is, as having values that are most different, positively or negatively, from the average values of all the goods in the choice set). BGS assume that a 'rational' consumer would maximise a linear utility function in which each attribute has a constant utility weight; by implication, these weights represent the consumer's true subjective valuations of the attributes. A non-rational consumer maximises a function in which the attribute weights are 'distorted' by salience, and so may 'undervalue' or 'overvalue' a good, depending on how its attributes compare with the corresponding averages. In our terminology, BGS are treating the rational utility function as representing the latent preferences of the non-rational consumer. The properties of the model ensure that, given sufficient observations of the choices of a non-rational consumer, her latent preferences can be recovered.

This methodology is developed with a more applied emphasis by Bleichrodt et al. (2001) and Li et al. (2014). These authors are primarily concerned with cases in

which a professional specialist has to make a decision in the best interests of a client. For example, consider a physician who has to choose between alternative medical treatments for an unconscious patient. The physician has access to data from a stated-preference survey in which the patient made various hypothetical choices between alternative probability distributions over health states. However, these responses are not fully consistent with one another, given the background assumption that 'the right normative model for decision under uncertainty' is expected utility theory — an assumption to which BPW are committed. According to BPW, such inconsistencies in stated-preference responses 'designate deficiencies in our measurement instruments that, even if the best currently available, do not tap perfectly into the clients' values'. The problems resulting from inconsistencies in stated preferences can be mitigated if those preferences are elicited in face-to-face interviews in which the client is asked to reconsider inconsistent choices (pp. 1498–1500, 1510). Notice the implicit assumption that the client has (or can be guided to form) preferences that are consistent with one another and with expected utility theory; the use of interviews is presented as a method of purifying preferences by the elimination of error.

But what if the physician has to make do with the patient's inconsistent survey responses? The real novelty of BPW's approach is their proposal of 'a quantitative manner for correcting biases in decision under risk and uncertainty when these cannot be avoided' (p. 1499). BPW use cumulative prospect theory (Tversky and Kahneman, 1992) as the *descriptive* model of choice while retaining expected utility theory as the *normative* model. There are two main differences between these models. First, cumulative prospect theory uses a *probability weighting function* to transform objective probabilities into their subjective counterparts; this transformation can be interpreted as taking account of psychological biases in the processing of probability information. Second, it has a *loss aversion* parameter which can be interpreted as picking up a bias induced by the framing of decision problems. Given these interpretations, an expected utility model of preferences can be constructed from an empirically estimated prospect theory model by replacing the estimated probability weighting and loss aversion parameters with the 'unbiased' values implied by expected utility theory. BPW propose that the patient's stated preferences should be used to estimate a prospect theory model, and that the 'corrected' expected utility model should be used to make choices on behalf of the patient. This is an econometric form of preference purification.

A somewhat similar methodology is proposed by Köszegi and Rabin (2007, 2008), who frame the problem as that of making inferences about individuals' preferences from observations of their choices while recognising that the reasoning that led to

these choice may have involved mistakes. Their main examples are of choice under uncertainty. Their approach is to infer an individual's subjective beliefs from the choices he makes between gambles with money outcomes, on the assumption that he prefers more money to less (contingent on any given state) and prefers higher probabilities of preferred outcomes to lower probabilities. If subjective beliefs, elicited in this way, do not coincide with objective relative frequencies, a 'revealed mistake in beliefs' is deemed to have occurred. The individual's preferences are then purified by working out what he would have chosen, had he acted on correct beliefs.

The work we have reviewed in this section can be understood as belonging to a common programme for reconciling normative economics with behavioural findings. This programme of *behavioural welfare economics* takes the objective of normative economics to be the measurement of the effects of economic policies on individual well-being, as assessed from the viewpoint of a social planner or entrusted professional (such as a physician, dietician or 'choice architect') who wishes to respect individuals' judgements about their own well-being. It treats cases in which an individual's choices depend on 'irrelevant' properties of framing as errors, 'error' being defined relative to the latent preferences that the individual would have revealed if not subject to reasoning imperfections. Latent preferences are assumed to satisfy conventional principles of rational consistency — in particular, context-independence. The satisfaction of latent preferences is taken as the normative criterion.

Implicit in this programme, as we understand it, is the idea that latent preference is a subjective concept. By this we mean that latent preferences are judgements or perceptions that are formed within the minds of individual human beings; they do not correspond directly with objective properties of the external world. Sunstein and Thaler appeal to this notion of subjectivity when they repeatedly insist that their aim is to make people better off *as judged by themselves*. Quite apart from issues of rhetorical strategy, there is a fundamental reason for thinking that the preference purification approach presupposes a subjective interpretation of latent preferences. Consider the implications of the contrary position, that behavioural welfare economics uses a criterion of (supposedly) objective well-being. What then would be the point of taking the circuitous route of reformulating that criterion as the satisfaction of latent preferences, defining a person's latent preferences in terms of the hypothetical choices that he would make in the absence of reasoning imperfections, and then postulating that such choices would maximise objective well-being? But if well-being is a subjective concept, preference purification can be defended as a

way of correcting errors in an individual's reasoning while respecting his judgements about his own well-being.

If the purification approach is to work, latent preferences must be coherent. But if latent preference is a subjective concept, the coherence properties that are attributed to them cannot be explained by the hypothesis that latent preferences map some objective concept that already has those properties. So the preference purification approach, as applied to any given individual, must presuppose that the individual has potential access to some mode of *latent reasoning* that generates subjective preferences that satisfy conventional principles of rational consistency. However, the writers we have considered so far do not explain what that mode of reasoning is, or how it generates coherent preferences. All they tell us is that it is free of the 'imperfections' that behavioural economists and cognitive psychologists have identified in actual human reasoning. This limitation of the preference purification approach perhaps stems from the structure of the standard theory of rational choice. That theory is formulated in terms of axioms of consistency among preferences, and between preferences and choices; it does not try to explain the reasoning by which individuals construct their preferences. In an analysis which uses this conceptual framework, latent reasoning is a black box. One of the interesting features of Hausman's defence of preference purification is that it looks inside this box.

## 1.4 Hausman on preference purification

Hausman (2012) discusses preference purification as part of a larger analysis of the economic concepts of preference and welfare. He says that this analysis 'clarifies and for the most part defends' the everyday practice of economics, while challenging some of the ideas that economists use when philosophizing about their work (p. i).

Hausman begins by trying to find a coherent interpretation of the concept of preference, as that is standardly used in positive economics. He proposes the following definition: 'To say that Jill prefers $x$ to $y$ is to say that when Jill has thought about everything she takes to bear on how much she values $x$ and $y$, Jill ranks $x$ above $y$' (p. 34). Thus, a preference is *comparative* ($x$ is *ranked* above $y$); the comparison is in terms of *value*; the valuation is *subjective* ('how much *she* values ...'); and it takes account of the *totality* of factors that the individual thinks relevant to the comparison ('*everything* she takes to bear on ...'). In short, a preference is a 'total subjective comparative evaluation'. Hausman claims that this definition 'matches

most of current practice' in economics, and urges economists to reserve the word 'preference' for this usage (pp. 34–35).

That the economic concept of preference is comparative and subjective seems uncontroversial. That it is also total is implicit in a fundamental feature of the role of preferences in economics — that preferences determine choices. (A person's choices can be influenced by any factors that she takes to be relevant. So if preferences determine choices, preferences must take account of any such factors too.) But Hausman's claim that preferences are *evaluations* is tied in with his reason-based understanding of choice. Since this claim turns out to be important for Hausman's defence of preference purification, we need to explain what he means by it.

Hausman interprets preferences as products of reasoning, and as premises that can be used in further reasoning about what to choose. It is, he says, a misconception to think that preferences are 'arbitrary matters of taste, not subject to rational consideration' (p. 18); they are 'more like judgments than feelings' (p. 135). He interprets the economic theory of choice as a theory of rational deliberation, in which individuals try to answer the question 'What do I have most reason to do?' (p. 5). He maintains that, in using a theory of rational choice, economics is committed to the claim that its explanations of an individual's choices are expressed in terms of reasons that *justify* those choices. Thus:

> [E]conomists regard ordinal utility theory as both a fragment of a positive theory that explains and predicts choices and as a fragment of a theory of rational choice that specifies conditions that preferences must satisfy in order to justify choices. (p. 20)

In arguing that this interpretation of 'preference' is faithful to the practice of economics, Hausman (pp. 19–20) considers how the axioms of choice theory might be justified as properties of sound reasoning about choice. He agrees with Broome (1991a) that the logic of total comparative evaluation requires transitivity. (If, all things considered, $x$ is more valuable than $y$ and $y$ is more valuable than $z$, then necessarily, $x$ is more valuable than $z$.) While not claiming that the completeness axiom is logically required in the same way, Hausman points out that if an individual's choices are always (i.e. given any feasible set) to be determined by preferences, preferences must be complete. Thus, completeness is 'a boundary condition on rational choice'. The implicit axiom that preferences are context-independent excludes 'factors that ought to be irrelevant' for total comparative evaluations. Thus, if a rational choice is one that is justified by sound and relevant reasons, and if reasons are to take the form of total comparative evaluations of the feasible options, rational choice is possible in general only if

those evaluations are complete, transitive and context-independent. In this sense, Hausman's interpretation of 'preference' offers an explanation of why the axioms of choice theory are treated as principles of *rationality*.

Having settled on a definition of 'preference', Hausman goes on to consider the role of preferences in welfare economics. He takes it as self-evident that welfare economics is concerned with individual well-being, assessed from some neutral viewpoint. He does not specify explicitly who is the addressee of welfare economics, but his intermittent references to 'legislators'. 'policy makers' and 'policy analysts' (e.g. pp. 89, 93, 95-97, 100-101) imply that, in the language of economics, the addressee he has in mind is a social planner who is seeking to promote well-being. In these respects, Hausman's approach to normative economics is aligned with that of behavioural welfare economics.

One of Hausman's central claims is that 'the satisfaction of preferences — even when preferences are informed, rational, and generally spruced-up — does not constitute well-being' (p. 77). Hausman sets out to show that preference-satisfaction theories of well-being — that is, theories that treat preference satisfaction as a (or even the only) constituent of well-being — are 'mistaken' and 'untenable' (pp. 86, 88). In line with his claim to defend the 'most part' of the practice of economics, Hausman allows that, in fact, satisfying preferences usually contributes to well-being. But this is because, in arriving at total evaluations of the options from which they can choose, individuals are in most cases strongly influenced by reasonable beliefs about how each option would affect their well-being. In such cases, preferences provide reliable *information about* well-being, but preference-satisfaction still does not *constitute* well-being.

However, Hausman (pp. 81–83) points to various cases in which, he claims, the satisfaction of preferences may *not* promote well-being. One such case is where individuals' evaluations of options are based on beliefs that are in fact false. Another is where individuals consciously choose to act contrary to self-interest. For our purposes, the most relevant case is where 'preferences are the result of ... problematic psychological mechanisms'. Following the practice of behavioural welfare economics, Hausman treats these mechanisms as inducing *mistakes*:

> Contemporary psychology has identified contexts in which people are likely to make mistakes, and policy analysts can use these findings to help decide whether to take people's preferences as guides to their welfare. One advantage of understanding that preferences are total comparative evaluations is that economists and regulators can make clear sense of

people's preferences being *mistaken*. (p. 100, italics in original)

Expanding on this idea in relation to cost-benefit analysis, Hausman says that in deciding whether to use preferences as indicators of well-being, one should ask whether the context in which the preferences are revealed is one in which preferences are 'undistorted'. Having defined the 'net benefit' of a policy of the excess of gainers' willingness to pay over losers' willingness to accept (p. 93), he says:

> The cost-benefit analyst should avoid relying on net benefits when preferences are distorted by decision-making flaws because the flaws provide a good reason to doubt that such preferences are a good guide to the individual's welfare. Examples of such flaws include overconfidence, exaggerated optimism, status quo bias, inertia, inattention, myopia, conformity, akrasia, and addiction. (p. 100)

So what *should* cost-benefit analysts rely on? Hausman's answer is that '[t]he best economists can do when they recognize flaws in people's deliberative capacities is to minimize their influence'. More specifically:

> [W]hen preferences are self-interested, well-informed, and undistorted ... it is sensible for those seeking to promote welfare to employ methods of appraising policies such as cost-benefit analysis that rely on information concerning preference satisfaction. When these conditions are not met, it makes sense to take steps to purify people's preferences of mistake and distortion so as to widen the domain in which these conditions are met and to attempt to measure expected benefit rather than preferences. (p. 102)

So Hausman's proposal for dealing with preference inconsistencies is essentially the same as that of behavioural welfare economics: preference purification.

Recall Hausman's argument that if preferences are to provide reasons for choice, they must be complete, transitive and context-independent. If a cost-benefit analyst is to use a person's purified preferences as indicators of that person's well-being in arriving at policy recommendations, then (at least within the relevant policy domain) purified preferences must have those same properties. So for Hausman's proposal to work, each individual's undistorted reasoning must generate latent preferences that are complete, transitive and context-independent. But can we expect this to be the

case?

In Hausman's analysis of rational preference formation, the agent is represented as engaging in sound reasoning about the truth or falsity of propositions. A preference is a particular kind of proposition — a total subjective comparative evaluation — that the agent holds to be true. This conceptual framework allows a definition of 'distorted' or 'mistaken' reasoning as reasoning that contravenes principles of conceptual coherence or valid inference, broadly understood. Since preference purification removes the effects of such reasoning, it will result in a set of preference propositions that are consistent with one another and with other propositions to which the agent assents. Since preferences may be derived from subjective propositions (for example, judgements about what is desirable or about the constituents of well-being), this account of preference formation preserves the subjectivity of preferences. As we have explained, Hausman is able to argue that sets of preference propositions that violate transitivity or context-independence are inconsistent and hence incapable of being generated by sound reasoning from consistent premises. But he explicitly denies that sound reasoning necessarily generates a preference relation that is complete (p. 19); all he can say is that completeness is necessary *if choices are to be determined by preferences*.

Thus, Hausman's analysis does not resolve the problem we identified in the literature of behavioural welfare economics. That problem was to justify the implicit assumption that, for any given individual, there exists some mode of latent reasoning that generates complete and context-independent subjective preferences. Were there such a mode of reasoning, it could be argued that context-dependent choices are the result of reasoning imperfections — that is, of failure to recognise the implications of sound reasoning. But Hausman's analysis leaves open the possibility that there are pairs of options for which sound reasoning is unable to determine a preference ranking. The implication is that context-dependent choices are not necessarily mistakes that can be corrected by purification.

## 1.5 The inner rational agent

As we noted in section 1.2, Hausman is less favourably disposed to nudge policies than are most contributors to behavioural welfare economics. His reservations about nudging, presented in a jointly-authored paper (Hausman and Welch (2010); henceforth 'HW'), throw light on the model of agency that underlies his analysis of preference purification.

In their opening summary of libertarian paternalism, HW express agreement with

many of Sunstein and Thaler's conclusions *about welfare*. For example, in relation to TS's argument for nudging people to save more for retirement, HW say that TS 'are impressed by the imperfections in individual decision-making illustrated by the extent to which people's choices to save for retirement are influenced by details concerning enrolment that ought to be of negligible importance'. and that TS 'catalogue many factors that can lead to mistakes in human judgment and decision-making'. For the purposes of their paper, HW do not need to defend TS's specific judgements about 'which factors interfere with rational deliberation'. but they endorse those judgements as 'generally plausible'.

HW's reservations about libertarian paternalism are not about its analysis of welfare, but about the nudging policies that it recommends. These reservations are formulated in terms of *autonomy*, defined as 'the control an individual has over his or her evaluations and choices'. If one is concerned about autonomy, HW say, 'there does seem to be something paternalistic, not merely beneficent, in designing policies so as to take advantage of people's psychological foibles for their own benefit' (p. 128). Throughout the paper, nudges are contrasted with 'rational persuasion'. For example:

> The reason why nudges such as setting defaults seem ... to be paternalist, is that in addition to or apart from rational persuasion, they may 'push' individuals to make one choice rather than another... [W]hen this 'pushing' does not take the form of rational persuasion, their autonomy — the extent to which they have control over their own evaluations and deliberations — is diminished. Their actions reflect the tactics of the choice architect rather than exclusively their own evaluation of alternatives. (p. 128)

And (having defined 'shaping' as 'the use of flaws in human decision-making to get individuals to choose one alternative rather than another' [p. 128]):

> [R]ational persuasion respects both individual liberty and the agent's control over her own decision-making, while, in contrast, deception, limiting what choices are available or shaping choices risks circumventing the individual's will. (p. 130)

But what do HW mean when they refer to 'the individual' or 'the agent' as an entity that may or may not have control over his or her evaluations, deliberations and choices? Notice that this agent is not a real human being whose thoughts and

actions are governed by psychological mechanisms. If the choices of the real human being are influenced by factors that cannot be construed as good reasons, HW are able to claim that this agent's will has been circumvented. The implication seems to be that the agent is capable of error-free autonomous reasoning that is undistorted by 'problematic' human psychological mechanisms. It is open to rational persuasion, but impervious to attempts to influence it by other means. Given any decision problem, it can identify the option that it wishes to choose, referring 'exclusively' to its own evaluations of alternatives. This seems to imply that the agent's reasoning can generate complete, transitive and context-independent total comparative evaluations. We will call this disembodied entity the *inner rational agent*.

Notice how ordinary human psychology is being treated as a set of forces that are liable to restrict the inner agent's ability to act according to the implications of its own reasoning. It is as if the inner rational agent is separated from the world in which it wants to act by a *psychological shell*. The human being's behaviour is determined by interactions between the autonomous reasoning of the inner agent and the psychological properties of the outer shell. However, in relation to issues of preference and judgement, the inner agent is the ultimate normative authority.

Something like this model of human agency seems to be implicit in Hausman's account of preference purification. Preference purification can be thought of as an attempt to reconstruct the preferences of the inner rational agent by abstracting from the distorting effects of — by 'seeing through' — the psychological shell. We suggest that behavioural welfare economics rests on a similar model of agency, albeit with a less fully-developed account of the reasoning of the inner rational agent. Recall that Sunstein and Thaler's criterion of well-being is given by the preferences that the relevant individual would reveal, were she to pay full attention to decision problems and to possess complete information, unlimited cognitive abilities, and complete self-control. One might think of these preferences of those of an inner rational agent whose reasoning is free of *internal* errors but which depends on faulty psychological machinery to provide it with information, to carry out complex information-processing operations, and to execute its decisions. Lack of attention can cause faults in the flow of information to the inner agent; limited cognitive ability can cause faults in information processing; lack of self-control can cause faults in decision execution. Preference purification is an attempt to reconstruct the decisions that the inner agent would execute if the faults in the psychological shell were corrected.

Before going further, let us clarify the status of our metaphor of the 'inner rational agent' trapped in a 'psychological shell'. We do not pretend that Hausman and BWE consider that the inner agent *actually* exists. In particular, we do not make any metaphysical claim about the existence of such an entity. It is possible that some behavioural welfare economists actually believe in the empirical relevance of this model, but the point we intend to criticise is that, *if* such an agent existed, then it would be possible to define unambiguously a single consistent and context-independent preference ordering, the individual's *true preferences*. Our criticism of the model of the inner rational agent is therefore a criticism of the notion of true preferences, defined as the counterfactual preferences the agent would hold if, instead of using some complex psychological processes, she was able to reason perfectly (i.e. to reason without being limited by her cognitive abilities, her imperfect information or her issues of self-control). The 'inner rational agent' corresponds to the counterfactual agent the individual would be under TS's conditions of unlimited cognitive abilities, perfect information and complete willpower: the true preferences of the individual are then defined as the revealed preferences of the inner rational agent.

## 1.6 Is the model of the inner rational agent tenable?

Let us begin by recording our surprise that so many behavioural economists have wanted to use the model of the inner rational agent. One of the first impulses for what is now called behavioural economics was a recognition that the mental processes that people actually use in decision-making do not necessarily generate choices with the rationality properties traditionally assumed in economics. An obvious corollary of this idea, pointed out by Kahneman (1996), is that rational choice is not self-explanatory: cases in which behaviour is consistent with the conventional theory of rational choice are just as much in need of psychological explanation as are deviations from that theory. The model of the inner rational agent seems to depend on a denial of this corollary. In that model, human psychology is represented as a set of forces which affect behaviour by *interfering with* rational choice, but rational choice itself — represented by the error-free reasoning of the inner agent — is not given any psychological explanation. Kahneman is right to say that this modelling strategy is 'deeply problematic' (pp. 251–252)[5].

[5]One might also see it as a methodologically questionable attempt to conserve the neoclassical theory of rational choice in the face of disconfirming evidence by re-interpreting it as applying, not to real human beings, but to imaginary disembodied agents. Compare Berg and Gigerenzer (2010) critique of 'as-if behavioural economics'. See also Goldstein and Gigerenzer (2002, p.75) criticism

It might be objected that the model of the inner rational agent has psychological foundations in dual-process theories of the mind. The idea that the workings of the mind can be separated into two 'systems' — the fast and automatic *System 1* and the slow and reflective *System 2* — has been suggested by a number of psychologists (e.g. Wason and Evans (1975), Kahneman (2003)), and is the central theme of Kahneman (2011)'s overview of his contributions to psychology and behavioural economics. Since Sunstein and Thaler use the same idea as an organising principle when reviewing behavioural findings (TS: 19–39), it is plausible to conjecture that they are thinking of the inner rational agent as System 2 and the psychological shell as System 1.

Recently, there has been something of a fashion for economists to appeal to dual-process neurological theories to motivate models of time-inconsistent behaviour. In these models, individual behaviour is determined by interactions between two neural systems — one far-sighted, rational and strategically sophisticated, the other either short-sighted and naïve or automatic. For example, Bernheim and Rangel (2004) explain the decision-making of drug addicts in terms of interactions between a 'cold' mode of reasoning, capable of solving dynamic stochastic programming problems, and a 'hot' mode of automatic responses to environmental cues. Benhabib and Bisin (2005) use a similar model of interaction between 'controlled' and 'automatic' processes to explain consumption and saving behaviour. Fudenberg and Levine (2006) and Brocas and Carrillo (2008) present models in which both systems are represented as rational maximisers, but one is far-sighted and the other is myopic. This dual-process modelling strategy is perhaps reasonable as a way of representing the decision-making of the type of drug user who repeatedly tries and fails to quit, or of the former drug user who consciously tries to avoid cues that might induce recidivism. One might be more sceptical when the same strategy is applied to everyday cases of preference inconsistency, as when Fudenberg and Levine explain the high levels of risk aversion observed in laboratory experiments by hypothesising that the typical student subject has deliberately constrained her own access to cash as a way of solving a self-control problem.

But even if one accepts the dual-process theory as a useful way of organising ideas about human psychology, the model of the inner rational agent remains vulnerable to Kahneman (1996) critique. One is not entitled simply to assume that the mental processes of System 2 can generate preferences and modes of strategic reasoning that are consistent with conventional decision and game theory. Indeed, that assumption does not fit easily with the logic of dual-process theory. One of the fundamental insights of that theory is that the automatic processing

of work in psychology that treats heuristics as 'poor surrogates for optimal procedures'.

mechanisms of System 1 are evolutionarily older than the conscious mechanisms of System 2. Thus, except in so far as its original features have atrophied, we should expect System 1 to be capable of generating reasonably coherent and successful actions without assistance from other processes. But if System 2 processes are later add-ons, there is no obvious reason to expect them to be able to work independently of the processes to which they have been added. Kahneman (2011, p.24) hints at the subsidiary role of System 2 when he says that '[w]hen System 1 runs into difficulty, it calls on System 2 to support more detailed and specific processing that may solve the problem of the moment'. The suggestion seems to be that, in dealing with choice problems, System 2's primary role is to provide decision support services. It is not obvious that this system always needs to be capable of making decisions in its own right, as the inner rational agent is supposed to be.

We have suggested that there can be choice problems that lack determinate rational solutions (with the implication that System 2 may not be able to solve them). To explore this possibility further, we return to the case of SuperReasoner in the cafeteria. Recall that SuperReasoner is a re-engineered version of an ordinary human being, Joe. He differs from Joe by not being subject to reasoning imperfections: he has no limitations of attention, information, cognitive ability or self-control. In all other respects, however, he is the same as Joe. According to Sunstein and Thaler, SuperReasoner's choices reveal Joe's latent preferences. Suppose that the options available at the cafeteria include cream cake and fruit. Were Joe to go to the cafeteria, he would choose (and would be willing to pay a small premium for) whichever of those two options was displayed more prominently. The cafeteria director has read *Nudge*, and wants to use the display that best satisfies Joe's latent preferences. Thus, she needs to know what SuperReasoner would choose. This raises the question that we asked (but did not answer) in section 1.3 Would the probability of his choosing cake be independent of the position of cake on the counter?

One way of answering this question builds on Bacharach (1993, 2006) analysis of frames. Joe's preferences are context-dependent because the problem of choosing between food items can be framed in different ways. Or, more accurately: Joe can *represent the problem to himself* in different ways. In one case, he construes the problem as a choice between (say) 'the cake at the front of the counter' and 'the apple at the back'; in another, he construes it as a choice between 'the apple at the front' and 'the cake at the back'. In the first case, he prefers 'the cake at the front'; in the second, he prefers 'the apple at the front'. Joe's preferences, *as viewed by Joe himself*, are not inconsistent. They are inconsistent only as viewed

by a theorist who conceptualises 'the cake (or apple) at the front' as the same thing as 'the cake (or apple) at the back'. As Bacharach (2006, p.13) puts it, whether a decision-theoretic principle of rationality has been violated 'depends on how *we*, the theorists, "cut up the world" '. But, he goes on: 'For decision theory, there are no unproblematically given "same things" '. If this is right, decision theory cannot legitimate the assumption that there is a single correct way of framing the cafeteria problem, accessible to any agent who, like SuperReasoner, is free of reasoning imperfections.

If SuperReasoner's rationality is interpreted in terms of conventional decision theory, as Sunstein and Thaler perhaps intend, Bacharach's argument implies that SuperReasoner's preferences can be context-dependent. However, that argument has less force against Hausman's account of reasoning. Recall that, in that account, sound reasoning can recognise that certain factors 'ought to be irrelevant' for total comparative evaluations. To see where this approach leads, let us stipulate that the relative position of the food items on the counter is such a factor. So SuperReasoner cannot say that, all things considered, the cake is more valuable than the fruit if the cake has the more prominent display, but less valuable in the opposite case. Thus, if we accept Hausman's definition of 'preference'. SuperReasoner cannot hold context-dependent *preferences* between fruit and cake. But, since his *feelings* are the same as Joe's, he feels an inclination to choose the cake in the first case, and to choose the fruit in the second. Were his Einstein-like powers of reasoning to lead him to the conclusion that the fruit was more valuable, his Gandhi-like powers of self-control would allow him to overcome any inclination to choose the cake. But let us suppose that, given the premises on which his reasoning operates, the relative value of cake and fruit is undetermined.

If, as we have argued, latent preference is a subjective concept, this supposition does not seem to imply any contradiction. It is true that, by virtue of his special powers, SuperReasoner can access all the information that is relevant for the choice between fruit and cake. For example, he knows all the respects in which eating fruit would improve his health, and all the respects in which eating cake would give him immediate enjoyment. If the uniquely correct choice could be determined by applying some well-defined algorithm to this multi-dimensional information, SuperReasoner would have the computational powers to solve the problem. But we know of no argument, either in behavioural economics or in the theory of rational choice, that would justify the assumption that such an algorithm exists.

So let us maintain our supposition: SuperReasoner cannot determine whether, all things considered, cake is more valuable than fruit or vice versa. In Hausman's

sense, he has no (strict or weak) preference between these options. Still, he feels an inclination to choose whichever of cake or fruit is more prominently displayed. What principle of sound reasoning would he contravene by acting on this inclination, just as Joe would? If the answer is 'None'. as we believe it is, then SuperReasoner's choices, like Joe's, can be context-dependent.

If instead we are to insist that SuperReasoner's choices must be context-independent, we seem to need to make completeness of preferences an *axiom* of reasoning, rather than a property that, depending on circumstances, reasoning may show to be true or false. Building on Hausman's characterisation of completeness as a boundary condition on rational choice, one might perhaps stipulate that, if an agent is to be truly rational, his choices must always be justified by preferences. One might then claim that rationality requires the agent to ensure that the set of preference propositions he holds to be true is sufficient to pick a nonempty set of justified choices from any nonempty set of options. Our own view (and, apparently, Hausman's) is that this requirement would be unwarranted; but let us set these reservations aside[6]. If SuperReasoner wanted to comply with the requirement, he would have to fill in the gaps in his otherwise incomplete preference ordering by constructing additional preferences whose content was not justified by reasoning.

But this conclusion is of no help to behavioural welfare economics. The problem that needs to be solved is that of discovering Joe's latent preference between fruit and cake. The line of argument we are exploring leads to the conclusion that, were Joe truly rational, he would have *some* context-independent preference between the two options. But that means only that the imaginary SuperReasoner would have responded to the demands of rationality by constructing such a preference, arbitrarily if necessary. We may have no way of discovering what that imaginary preference would be. And it is only in the most tenuous sense that this imaginary preference is latent *in Joe*.

Savage (1954, pp.101-103) discussion of the Allais Paradox nicely illustrates the issues involved here. Savage reports that, when he was first presented with Allais' two choice problems, he expressed a preference for Gamble 1 in Situation 1 and for Gamble 4 in Situation 2 — the response that constitutes the Paradox and that is

[6]Some readers may be tempted to think that this requirement could be justified by a 'money pump' (or 'Dutch book') argument, but that thought would be mistaken. Invulnerability to money pumps does *not* imply that choices are determined by preferences (Cubitt and Sugden, 2001). As a simple counter-example, consider an agent who acts on the decision heuristic of never making exchanges, whatever her initial endowment and whatever options are available to her. This heuristic implies a pattern of choice that cannot be rationalised by any (reference-independent) preference relation, but which is clearly invulnerable to money pumps.

inconsistent with Savage's own expected-utility axioms (one of which is an axiom of completeness). He confesses that he 'still feel[s] an intuitive attraction to those preferences'. However, since his analysis of expected utility is intended as a normative theory, it would be an 'intolerable discrepancy' for him to maintain two preferences that together were inconsistent with the axioms of the theory:

> In general, a person who has tentatively accepted a normative theory must conscientiously study situations in which the theory seems to lead him astray; he must decide for each by reflection — deduction will typically be of little relevance — whether to retain his initial impression of the situation or to accept the implications of the theory for it. (p. 102)

Savage reassures himself of the validity of his axioms by re-framing the four gambles so that their outcomes all depend on the same draw from a set of lottery tickets numbered 1–100. Prizes are assigned to tickets so that the prizes for Gambles 1, 2, 3 and 4 respectively, in units of \$100,000, are (5, 0, 5, 0) for ticket 1, (5, 25, 5, 25) for tickets 2–11, and (5, 5, 0, 0) for tickets 12–100[7]. Since, in each situation, the two gambles on offer differ only in the event that one of tickets 1–11 is drawn, Savage concludes that the other tickets are irrelevant to the decisions that have to be made. Conditional on this event, Gambles 1 and 3 are identical, as are Gambles 2 and 4. Thus, Savage's original preferences are unacceptably context-dependent. Both of them cannot be right. But which of them is wrong? Savage tells himself that, in both situations, the choice problem reduces to 'whether I would sell an outright gift of \$500,000 for a 10-to-1 chance to win \$2,500,000'. Consulting his 'purely personal taste'. he finds that he prefers the former. He then accepts the implication that he prefers Gamble 3 to Gamble 4, saying: 'It seems to me that in reversing my preferences between Gambles 3 and 4 I have corrected an error'.
Notice that Savage has invoked a third situation (let us call it 'Situation 3') in which he has to choose *either* \$500,000 with probability 1 ('Gamble 5') *or* \$2,500,000 with probability 10/11 and nothing with probability 1/11 ('Gamble 6'). According to his axioms, his ranking of Gamble 3 relative to Gamble 4 (and, equivalently, his ranking of Gamble 1 relative to Gamble 2), should be the same as his ranking of Gamble 5 relative to Gamble 6. He finds an inclination to prefer Gamble 5 to Gamble 6. So far, this is not a resolution of the original problem; it is merely an expansion of the set of inconsistent preferences. However, it seems that Savage feels more confident about his inclinations in Situation 3 than in the other two situations, and so decides to use

[7]The implicit claim that the original and revised versions of the problems are equivalent to one another is open to question, but it is an implication of Savage's axioms.

those inclinations as his arbiter. There is nothing wrong with that: as Savage says, this is a matter of reflection, not deduction. But there seems no reason to suppose that this particular sequence of reflections leads to the uniquely correct resolution of the original inconsistency (if inconsistency it is). At most, this story tells us that if someone genuinely accepted the expected-utility axioms as requirements of rationality and was not cognitively constrained, he would be able to settle on *some* preferences, consistent with those axioms, which he was willing to live with (but which might still be contrary to his actual inclinations). That is not particularly helpful if we are trying to identify the actual latent preferences of an ordinary Joe whose choices and inclinations have the Allais Paradox pattern.

## 1.7 Purification — or regularisation?

We have argued that latent preference is not a useful concept for normative economics. How then (a sceptical reader might ask) can we explain the fact that so many behavioural economists have wanted to use it? We suggest that this practice may be a by-product of a modelling strategy that is common in behavioural economics. When used in the development of descriptive theories, this strategy has significant methodological virtues, but it is liable to lead one astray in normative work.

This strategy, which we will call *behavioural optimisation*, uses conventional rational-choice theory as a template, and models the individual as maximising a *behavioural utility* function that retains many of the properties of the utility functions used in neoclassical economics and game theory. Psychological factors that are neglected in conventional theory are modelled by allowing behavioural utility to depend on additional variables. Often, the standard utility function is represented as a special case of the behavioural function. Failing that, the two functions can usually be thought of as distinct special cases of a more general utility function.

If the objective is to develop a parsimonious descriptive theory that generates successful predictions about economic behaviour, this strategy has obvious practical merits. If one accepts (as many behavioural economists do) that the predictions of conventional economic theories are often good first approximations to the truth, it may be more productive to look for incremental improvements to those theories than to start again from scratch and to re-invent wheels. Even if one is sceptical about the predictive success of conventional theory, it remains true that economists have developed a large body of abstract theoretical results that hold for maximising

behaviour in general, and which can be re-used in behavioural utility models. Using behavioural utility functions also makes it easier to identify and test the novel implications of behavioural theories, and to measure the increase in explanatory power that can be attributed to the inclusion of additional variables. Exactly these arguments are used by Rabin (2013) to defend the behavioural optimisation strategy. Similarly, Hausman (2012, pp.114-115) favours the strategy of modelling psychological factors such as framing through their effects on preferences on the grounds that 'if economists and decision theorists continue to regard preferences as determinative [of choices], then they can still employ consequentialist and game-theoretic models and the mathematical tools that permit predictions to be derived from them'[8].

Notice that for the behavioural optimisation strategy to have these merits, it is not necessary that the standard theory is a representation of *rational* choice; what matters is that it makes at least moderately accurate predictions across a wide domain. However, because the standard theory is usually interpreted in terms of rationality, it is tempting to think that this modelling strategy allows one to isolate the effects of mistakes (i.e. those effects on utility that occur because 'behavioural' variables do not take the values that correspond with the standard theory), and so to identify latent preferences (i.e. the preferences that would result if behavioural variables took their standard values). Rabin (2013, p.529) presents this feature of behavioural utility models as an important merit, on the grounds that it allows us to 'capture many errors in terms of systematic mistakes in the proximate value function ... or where the beliefs imported into their maximizations are systematically distorted'. The use of behavioural optimisation models to purify preferences was discussed in section 1.3, where it was exemplified by the work of Bleichrodt et al. (2001), Köszegi and Rabin (2007, 2008), and Bordalo et al. (2013).

We will argue that this method of defining and identifying latent preferences is unsatisfactory. We will develop this argument by considering how BPW's purification methodology might be applied to Savage's version of the Allais Paradox.

Recall that BPW use cumulative prospect theory as the descriptive model of choice. Viewed in the perspective of that theory, Allais' four gambles can be differentiated in terms of two characteristics — the probability of winning at least $500,000,

[8] In saying this, Hausman is in danger of undercutting his earlier argument that, in the practice of economics, preferences are implicitly interpreted as total subjective comparative evaluations (see section 1.4 above). If there are pragmatic reasons for behavioural economists to use preference-based models when explaining non-rational choices, neoclassical economists might favour preference-based models for pragmatic reasons too.

and the probability of winning $2,500,000. In terms of the second characteristic, Situations 1 and 2 are equivalent to one another. (In each situation, the probability of winning $2,500,000 is either zero or 0.10, depending on whether the first or the second gamble is chosen.) So an explanation of the Allais Paradox must work through the first characteristic. The probability of winning at least $500,000 is 1.00 in Gamble 1, 0.99 in Gamble 2, 0.11 in Gamble 3, and 0.10 in Gamble 4. The difference between the two relevant probabilities (1.00 and 0.99 in Situation 1, 0.11 and 0.10 in Situation 2) is the same in both situations, which is another way of explaining why the Paradox contravenes expected utility theory. However, cumulative prospect theory transforms each objective probability $p$ into a subjective decision weight $w(p)$. The Allais Paradox is possible if $w(1.00) - w(0.99)$ is sufficiently greater than $w(0.11) - w(0.10)$. That inequality is consistent with most empirical estimates of the probability weighting function, and also with intuition: the difference between the certainty of a very large prize and a 99 per cent chance of it *feels* more significant than the difference between an 11 per cent chance and a 10 per cent chance. So it is plausible to suppose that cumulative prospect theory is picking up a psychological mechanism that contributes *in some way* to the Allais Paradox.

BPW's purification methodology treats the non-linearity of the probability weighting function as a reasoning error that needs to be corrected if we are to identify latent preferences. But where is the error? Of course, there would have been an error *if* the decision-maker had known the utility to him of the three possible outcomes, *and if*, believing expected utility theory to be the right normative model, he had tried to calculate the expected utility of each of the four gambles, *and if* in doing so he had used decision weights in the mistaken belief that they were objective probabilities. But that is not a remotely plausible account of the reasoning that leads real people to choose Gambles 1 and 4. To point to just one problem with this account, remember that when people respond to Allais' problems, they are *told* all the relevant objective probabilities. If you were to ask a respondent what he believed to be the percentage probability of an outcome that he had just been told had a probability of 1 per cent, what answer would you expect to get?

What BRW's purification methodology reveals is that, *relative to the benchmark of expected utility theory*, the person who has made the Allais Paradox choices has behaved as if he held false beliefs about the probabilities of the relevant events. If expected utility theory could be interpreted as a first approximation to a true description of how people actually reason, it might be plausible to move from that as-if proposition to the conjecture that the person's actual reasoning followed the logic of expected-utility reasoning but with false beliefs. But the truth is surely

that expected utility theory provides a first approximation to *the choices that people actually make*, not to the reasoning by which they arrive at those choices.
It is not surprising that expected utility theory has this approximation property, at least when applied to lotteries with monetary outcomes and explicit objective probabilities. Whatever mental processes people use in decision-making about such lotteries, one would expect larger money prizes to be perceived more favourably than smaller ones, other things being equal. Similarly, for any given money amount $x$, one would expect larger probabilities of winning at least $x$ to be perceived more favourably than smaller probabilities. By generalising these two intuitions and by organising them in a simple and tractable functional form, expected utility theory picks up some of the main patterns in the decisions that are generated by actual human reasoning. In the case of the Allais Paradox, however, cumulative prospect theory provides a more accurate description of actual decisions. In the absence of a theory of how people reason, that is just about all that can be said. One is certainly not entitled to infer that Allais Paradox choices reveal errors of reasoning that are not committed by people whose choices are consistent with expected utility theory. We conclude that BPW's methodology does not reconstruct latent preferences. In fairness, however, it should be acknowledged that BPW sometimes justify this methodology in more pragmatic terms, as when they say:

> We are well aware that many of the assumptions underlying our proposal are controversial, such as the very existence of true underlying preferences. These assumptions are, however, the best that we can think of in the current state of the art for situations where decisions have to be taken, as good as possible, on the basis of quick and dirty data. (p. 1500)

Recall that BPW's paradigm decision problem is that of a professional specialist who has to make a choice on behalf of a client, given only partial information about the client's revealed or stated preferences. When BPW say that expected utility theory is 'the right normative model for choice under uncertainty' (pp. 1498–1499), they seem to be referring to the decision problem faced by the professional. One might perhaps argue that, if the professional shares BPW's view of the normative status of expected utility theory, she ought to construct *her* judgements about the client's welfare, and hence about the decisions *she* should make when acting on the client's behalf, so that these judgements are consistent with the expected-utility axioms. Viewed in this way, what seems to be required is not an inference about the hypothetical choices of the client's inner rational agent, but rather a way of *regularising* the available data about the client's preferences so that it is compatible

with the particular model of decision-making that the professional wants to use. Regularisation in this sense is almost always needed when a theoretical model comes into contact with real data. For example, consider an economic model of the spatial distribution of unemployment. Suppose that, in this model, every job-seeker and every job offer has a spatial location. In the world of the model, this makes perfectly good sense: each job-seeker has a 'home' and each job has a 'workplace'. But if we try to apply the model in practice, we will find that 'home' and 'workplace' can be ambiguous concepts. Some people have two or more home addresses, while some have none; and analogously for jobs. In order to regularise the data so that they fit the model's categorisation scheme, some more or less arbitrary classifications will need to be made. But one would surely not claim that these classifications correspond with latent truths about the world that real job-seekers and real employers have failed to recognise. In the same way, a medical decision-maker might reasonably use BPW's methodology to construct a tractable *model* of the client's preferences, regularised so as to be consistent with expected utility theory, without claiming that the preferences in the model were latent in the client. The arguments we have developed in this chapter would not be objections to a version of behavioural welfare economics that claimed only to regularise revealed preferences that were inconsistent with conventional theory, without interpreting this process as the identification and correction of errors, or as a way of helping individuals to make better choices. But that is not the version of behavioural welfare economics that is to be found in the literature.

## 1.8 Conclusion

In arguing for libertarian paternalism, Thaler and Sunstein (2008, p.6) criticise conventional economists for assuming that ordinary human beings are 'Econs' — an imaginary species which 'thinks and chooses unfailingly well'. Sunstein and Thaler claim that their own approach to behavioural welfare economics — an approach that is becoming part of the mainstream of behavioural economics, and whose core features are endorsed by Hausman (2012) — breaks away from this mistaken assumption, and models human psychology as it really is. We have argued that this claim is misleading. It would be closer to the truth to say that behavioural welfare economics models human beings as faulty Econs. Its implicit model of human decision-making is that of a neoclassically rational inner agent, trapped inside and constrained by an outer psychological shell. Normative analysis is understood as an attempt to reconstruct and respect the preferences of the imagined inner Econ.

We maintain that if behavioural and normative economics are to be satisfactorily reconciled, the first essential is that economists learn to live with the facts of human psychology. We need a normative economics that does not presuppose a kind of rational human agency for which there is no known psychological foundation. One possible line of advance is to find a normative criterion that respects individuals' choices without referring to the preferences — consistent or inconsistent — that lie behind them. Sugden (2004) 'opportunity criterion' is an example of this strategy. Such a criterion may seem unappealing if one presupposes that normative economics is addressed to a benevolent social planner, but this addressee is no more than a theoretical or literary construct. If one thinks of individual citizens as principals and public decision-makers as their agents, it may seem more natural to treat *citizens* as the addressees of normative economics. Citizens who recognise that their choices are sometimes context-dependent might still want their agents to respect those choices (Sugden (2013), see also chapter 4).

To readers who would prefer to conserve more of the framework of conventional welfare economics, we commend the 'regularisation' perspective that we sketched in section 1.7. Instead of claiming to reconstruct the latent neoclassical preferences of individuals whose psychology has led them into error, economists might think of themselves as doing their best to represent the complex reality of human judgement and decision-making in a highly simplified but perhaps still useful normative modelling framework. But that would require a significant retreat from the ambition — not to say hubris — of much of the current literature of behavioural welfare economics.

Chapter 2

# How we Became Rational: the Duality of the Economic Man

## Contents




**Abstract:** This chapter investigates the historical origins of the model of the inner rational agent. We suggest that the two main arguments justifying the notion of true preferences are grounded on two quite different conceptions of the 'economic man', one as the simplification of a real individual and another as a representative agent. We show that this duality finds its origins within the marginalist revolution: while Jevons and Menger considered economics as a science of individual choice, Walras and later Marshall considered it as a science of social institutions. These two specific methodological approaches generated two distinct figures of the economic man, and we suggest that the current ambiguity about the nature of the neoclassical economic man results from the homogenisation of those two approaches through Pareto's definition of the *Homo economicus*. We are then able to question the relevance of assuming that individuals present true preferences, since the coherence of those

preferences is probably a property of a specific social institution — repeated markets — rather than of individual choice.

## 2.1 Introduction

According to neoclassical economics, human behaviour can be modelled thanks to the rational choice theory, a purely economic theory of behaviour, separate from psychology and sociology, considering the behaviour of perfectly rational agents, who know what their objectives are and how to achieve them. Throughout this thesis, we will use the generic term 'rational choice theory' to refer to all the specific economic theories (e.g. decision theory, expected utility theory) whose primitive is the existence of coherent preferences, explaining behaviour as the result of acting in an instrumentally rational way (Hands, 2010). Economists are then continuously torn between two opposite interpretations of this 'economic man': on the one hand to interpret it as a fictive and axiomatic entity designed for investigation purposes — such as the study of market equilibrium — and on the other hand to interpret it as the representation of a real individual. We can indeed find a large literature trying to explain a wide range of human actions thanks to the rational choice theory, such as marriage (Becker, 1974), criminal activities (Becker, 1968) and even obesity and food behaviours (Ruhm, 2012). However, it is not clear whether the theory describes the behaviour of those agents under a set of ideal conditions (offering a somewhat realistic account of actual behaviour, and treating those coherent preferences as the counterfactual preferences of the individual, if she was perfectly rational) or simply that agents behave as if the theory was real (in a more fictionalist sense, Mäki (2003, p.501)). The crucial distinction here is that, in the former case, all those activities would effectively result from a conscious deliberation and an assessment of the costs and benefits of each option, whereas in the latter, we simply try to rationalise *a posteriori* individual behaviours. Although such fictionalist approach may be defended in the case of positive analysis, we may question its relevance when producing normative assessments: designing public policies indeed requires predicting their consequences, and therefore identifying the actual determinants of individual behaviour. Building public policies on an inadequate model of behaviour can indeed lead to counter-productive outcomes and crowding out effects (such as remunerating the blood donation, Titmuss (1970)). Throughout the rest of this work, we will always consider that we aim at producing economic theories of individual behaviour that allows for policy recommendations, requiring hence a relatively realistic explanation of individual behaviours.

If we consider that economic theory should produce realistic descriptive models of behaviour, then rational choice theory should probably be revised. Behavioural economists have indeed suggested abandoning the model of a rational *Homo economicus* as a descriptive model of human behaviour in order to produce more accurate and psychology based models of individual behaviour — an extreme form of this approach is neuroeconomics and the study of the neurological basis of economic behaviour (Loewenstein et al., 2008), but such models probably lose in simplicity and tractability what they may gain in realism.

Two main arguments have then been suggested in order to defend rational choice theory and its model of the economic man, but it appears that they are grounded on the two distinct interpretations we mentioned above. Levitt and List (2008) provide a first argument by highlighting a difference between experimentations in the laboratory on isolated individuals and the 'settings of interests' of economics:

> Perhaps the greatest challenge facing behavioral economics is demonstrating its applicability in the real world. In nearly every instance, the strongest empirical evidence in favor of behavioral anomalies emerges from the lab. Yet, there are many reasons to suspect that these laboratory findings might fail to generalize to real markets. [...]
> To be empirically relevant, the anomalies that arise so frequently and powerfully in the laboratory must also manifest themselves in naturally occurring settings of interest. Further exploring how markets and market experience influence behavior represents an important line of future inquiry. (Levitt and List, 2008, p.910)

Since the subjects of an experiment in the laboratory are not in the same conditions than real agents in a market, the individuals will not necessarily behave in the same way depending on whether they are only the subjects of an experiment or real actors in a market. This defence of the neoclassical model, suggested among others by Binmore, considers that experimental findings can only be considered as relevant — and therefore able to question rational choice theory — if the problem 'seems simple to the subjects', the 'incentives provided are "adequate" ' and the 'time allowed for trial-and-error adjustment is "sufficient" ' (Binmore, 1999, p.F17). Binmore indeed stresses that testing economic theory outside this specific framework is not relevant, since '[j]ust as we need to use clean test tubes in chemistry experiments, so we need to get the laboratory conditions right when testing economic theory' (Binmore, 1999, p.F17). There is within this first argument

the idea that the actual individuals progressively learn to behave rationally: the rational choice theory therefore describes the behaviour of real individuals who have achieved this state of rationality. Behavioural findings are therefore of little interest, since they correspond to deviations from the rational behaviour, and therefore only describe the mistakes the individual can make during this process of 'rationalisation'. They do not question the underlying rational behaviour described by the rational choice theory. From a normative perspective, this means that those deviations from the theory must be interpreted as *mistakes* (that people would not have done were they not boundedly rational): this claim can then be used to justify paternalistic interventions, since a social planner may identify what mistakes the individuals tend to do, and then help them to make better choices for themselves.

The second defence of the neoclassical model of the economic man is put forward by Levine:

> More to the point — it is crucial to recognize that the goals of psychologists and economists are different, and that this has implications for importing ideas from psychology into economics. The key difference between psychologists and economists is that psychologists are interested in individual behavior while economists are interested in explaining the results of groups of people interacting. [...] The need to study groups of potentially large numbers of people — as I write this we are approaching seven billions — imposes constraints on economic models of individual decision making that are not present for psychologists. Economists need simple and broad models of behavior. Narrow complex models of behavior — neurally-based descriptions, for example — cannot easily be used to study the behavior of many people interacting. Hence the focus by economists on axiomatic models that provide a reasonable description of particular data while also giving decent results over a broad range of social settings. (Levine, 2012, pp.125-127)

Levine argues that economists are not concerned with the effective behaviour of isolated individuals, but with the results of human interactions at the scale of the society. It is therefore not necessary to describe the behaviour of a real individual, but of an ideal one, representative of a group of individuals. Economists need then to define a set of axioms that can describe the behaviour of this ideal entity: they generally assume coherent preferences and self-interested motives, and justify it by appealing to evolutionary considerations (see for instance Alchian (1950) and

Friedman (1953) in the case of firms' decisions in repeated markets). We can also find this kind of argument in the first defence of rational choice theory: there is indeed the claim that the individuals tend to become rational, since they are able to adapt themselves during the repetition of the game. Referring to Lincoln's famous statement that '[y]ou may fool all the people some of the time, you can even fool some of the people all the time, but you cannot fool all of the people all the time', Levine argues that 'most people are rational most of the time' (Levine, 2012, p.76): there obviously are individuals who will not be (either some or all the time) rational, but in general, the individuals will behave rationally in society.
Levine's claim is that the *Homo economicus* does not describe the behaviour of an isolated individual — who can be rational or not — but of a representative individual: since most people are rational most of the time, we can approximate the behaviour of a given individual by the behaviour of an 'average' one, who is rational. The model of the *Homo economicus* is therefore relevant because the settings of interest of economics are social institutions, characterized by the interaction of a large number of individuals. Unlike the first argument according to which the actual individuals progressively learn to behave rationally, there is within this second approach the idea that behavioural findings are not particularly relevant towards the issues economists deal with: a model of a representative individual is therefore sufficient for economic purposes.

We have here the perfect illustration of the ambiguous interpretation of the rational economic man: either a real individual in specific settings or a fictive entity designed for the study of the interaction of large number of individuals. Quite interestingly, the authors of both approaches refer to the same kind of argument in order to justify the assumption of rationality: the evolutionary tendency of becoming more rational as long as the game is repeated. However, Binmore's claim is that *each* individual tends to become rational, whereas it would be sufficient in an institutional conception of economics to claim that the *representative* individual tends to become rational, i.e. that the social outcome can be described as the result of the interaction of rational agents. Although it may be possible that each individual tends to become rational in specific kinds of social interactions, such as competitive markets or specific experimental settings, it is not certain that a social outcome which can be represented by the interaction of individually rational agents is effectively the result of the interaction of individually rational agents. However, Levine's argument seems to be grounded on the identity between those two figures of the economic man as a real agent and as a representative one: he indeed refers to the rationality of the individuals in specific settings to justify

the rationality of the representative individual, although the rationality of the representative individual should not be deduced from the behaviour of isolated individuals, but from the structure itself of social interactions between a large number of individuals[1]. If the axiomatic characterisation of the representative economic man were deduced from the behaviour of real individuals, then economists could not use a model of a rational representative economic man when they model social institutions for which the individuals are not individually rational. We may now notice that, from a normative perspective, knowing whether it is the actual individual or the representative one who becomes rational is of great importance. Recommendations such as libertarian paternalism indeed relies on the assumption that people make mistakes, i.e. that there is an individual deviation from the right (and rational) behaviour: there are no such considerations when claiming that the representative individual becomes rational, since it does not claim that the underlying behaviours are effectively mistakes or not. In the former case, we consider that the revealed preferences of the individual differ from her true, coherent preferences, and therefore that the individual is likely to make mistakes due to cognitive limitations. In the latter, there is no value judgment about what is the right behaviour at the individual level, but at the collective one: it is indeed not the *individual* preferences that may be incoherent or not, but the *social* outcome.

We can therefore notice that the defences of rational choice theory are grounded on the ambiguity of the nature of the neoclassical economic man — as the model of a real individual in specific settings, or as a fictive and representative entity. Indeed, the claim that individuals tend to become rational and can therefore be modelled by a rational representative agent does not support the extension of the scope of economics to the study of human behaviours for which the agents are not individually rational. Furthermore, both interpretations lead to quite different normative recommendations, since a departure from the recommendations of rational choice theory will be interpreted as an individual mistake in the first case and as an 'institutional' mistake in the second case.

[1] Levine takes the example of the 'rush hour game' (Levine, 2012, pp.12-13) — in which each driver chooses a road in order to minimize her time of travel — in order to highlight that the individuals are generally rational. Since — from Levine's own experience — it is apparently not possible to gain time by choosing an alternative road, he concludes that the situation is a Nash equilibrium and argues that it results from the individual rationality of each driver, who used to try alternative roads without success. The rationality of the drivers is however deduced from the property of the whole traffic: it means that the representative driver is rational, but not necessarily each driver.

In this chapter, we suggest investigating the origins of this ambiguous definition of the economic man, so as to properly understand the nature of the hypothesis of coherent preferences in economics. We show that this hypothesis probably results from a specific property of repeated markets rather than of individual choice: this seriously questions the validity of new behavioural economics (Davis, 2011, chap. 2) and of its different developments in normative economics such as libertarian paternalism, since there is less support for assuming that people may present underlying coherent preferences. We will start our investigation by studying the origins of microeconomics — which required a model of economic man — and of the neoclassical school in the 19th century. We highlight that the marginalist revolution produced two different conceptions of economics, one as a science of human action, and the other as a science of human activities. It resulted from this duality two different models of the economic man, one as the idealisation of a *real* individual, and the other as the definition of a *fictive* individual, defined from the properties of a social phenomenon. We then show that those two economic men had been homogenized through the figure of the *Homo economicus* developed by Pareto: although it was fundamentally designed for the study of human institutions (repeated markets) and not human actions, Pareto tried to ground his model of representative agent on individual choice properties and integrated within the individual a 'rational dimension' — which appeared to be a property of repeated markets rather than of individual choice.

## 2.2 Marginalism and the development of microeconomics

In the 1870's, Jevons, Menger and Walras almost simultaneously and independently published their main work[2], all of them grounded on a subjective theory of value. As underlined by Jaffé (1976), the historiographical practice then 'homogenized' their thoughts, by considering them as the independent discoverers of the marginal utility principle, without paying a closer attention to the differences of their respective approaches. We highlight in this section some of the main differences of their works, and in particular their distinct conception of the economic man.

[2] Jevons (1871), *Theory of Political Economy*, Menger (1871), *Grundsätze der Volkswirthschaftslehre*, and Walras (1874), *Éléments d'économie politique pure, ou Théorie de la richesse sociale.*

### 2.2.1 The origins of microeconomics and the use of mathematics

The main theoretical contribution of the marginalists — by opposition to the classical economists — is the explanation of the value of a good by its exchange value, defined as the ratio of the marginal utilities of the consumption of each good, rather than in terms of labour or any objective measure. The marginalists therefore produced a subjective theory of value, according to which the value of a good is determined by the personal tastes of the individuals, and focused on the phenomenon of the exchange in order to explain the formation of prices and the value of goods. As argued by Hébert (1998), it is probably this shift from macroeconomics to microeconomics that really constituted the paradigm shift between classical and neoclassical economics, more than the growing place of the notion of utility in economics as a motivating factor in human behaviour. His argument is that the need for a marginal notion of utility appeared during the industrial revolution, with the development of large infrastructures — such as the railroad —, whose marginal cost was different from the average cost: the real pioneers of the marginalist analysis were therefore engineers, such as Ellet (1839), Dupuit (1844, 1849), and Lardner (1850).
Ekelund and Hébert (1999) suggest that the origins of modern microeconomics can be traced back to the works of those engineers — more particularly the French engineers from the Corps des Ingénieurs des Ponts et Chaussées — and that Jevons, Menger and Walras had all been influenced by their works. Jevons for instance stated that:

> To Lardner's Railway Economy I was probably most indebted, having been well acquainted with that work since the year 1857. Lardner's book has always struck me as containing a very able investigation, the scientific value of which has not been sufficiently estimated. (Jevons, 1871, p. xviii)

The case of Walras is quite peculiar, since he was very critical of the work of those engineers — and more particularly Dupuit's work — and only recognized his father Auguste Walras and Cournot as sources of inspiration for his work[3]. It seems however that his position was much more due to personal and affective reasons than scientific ones: Cournot was indeed a former classmate of his father Auguste Walras at École Normale Supérieure, and he confessed in his autobiography that

[3]In a letter addressed to Jevons (23/05/1874), Walras explicitly stated that the only preceding works that helped him were the ones of his father and of Cournot (Jaffé, 1965, p.397). The different letters from and to Walras quoted in this article can all be found in the first volume of Jaffé (1965): we will therefore only mention the pages of the different letters throughout this chapter.

he had no appetite for the technical details of engineering (Jaffé, 1965, vol.I, p.2). He also failed several times to graduate from engineering schools such as École Polytechnique and École des Mines. Those different elements could explain a certain resentment towards French engineers, and the systematic criticism of their work: Walras for instance blamed Dupuit for confusing the utility curve (i.e. the marginal utility an individual gets from the consumption of the good) and the demand curve (the willingness to pay for the good), since the willingness to pay a good also depends on the personal wealth, as well as the utility we get from the other goods[4]. We can however find a similar and earlier critic with Bordas (1847). Dupuit defended his approach by referring to a partial equilibrium analysis: he indeed recognized the influence of the personal wealth as well as the influence of the prices of other goods on the demand, but considered that economists could also focus on the analysis of a single market, since those different elements do not prevent the existence of a price for each object, each person at each moment (Dupuit, 1849, p.184). It seems therefore that Walras criticized Dupuit's work by referring to general equilibrium considerations, whereas Dupuit explicitly placed his analysis in a partial equilibrium framework.
Furthermore, and unlike Jevons and Menger, the notion of marginal utility is relatively secondary in Walras' work: his objective was indeed to build a theory of price in interconnected markets, and the idea of marginal utility was only a convenient tool to deduce the demand curve from the utility curve (he did not in particular intend to elaborate a theory of subjective value in consumption, Jaffé (1976, pp.514-515)).

In the case of Menger, no obvious link to this engineering tradition exists, since, according to Hayek, he was not acquainted with their works (Menger, 1871, pp.14-15) and he also did not explicitly refer to them in the *Grundsätze*[5]. This could partly explain one of the most visible differences between Jevons, Menger and Walras, the use of mathematics: the mathematical tools in economics introduced for instance by Cournot and developed by the engineers are indeed of a great importance in the work of Jevons and Walras, but are lacking in Menger's theory.

[4]This criticism is developed in the 41st lesson of the *Éléments d'économie politique pure* (§385 to 387) as well as in his correspondence with Jevons (28/02/1877, p.533; 25/05/1877, p.535). Quite interestingly, Jevons recognized that 'Dupuit had a very profound comprehension of the subject', unlike Walras who considered that Dupuit's work had little merit.

[5]Hébert (1998) nevertheless argues that Menger's personal library contained several french economic journals — such as the Annales des Ponts et Chaussées, the Journal des économistes and the Journal de la statistique de Paris, in which we can find for instance Dupuit's work — and was therefore aware of this engineering tradition.

However, the main argument of Menger against the use of mathematics in economics was probably on a methodological level. Firstly, his criticism was directed toward differential analysis, since it requires the specification of continuous utility and demand functions, which are quite unrealistic[6]. He also expressed another criticism in his correspondence with Walras:

> Allerdings gehöre ich nicht zu des eigentlich Anhängern des mathematischen Methode der Behandlung unserer Wissenschaft. Ich bin nämlich der Meinung, dass die mathematische Methode der Hauptsache nach eine solche des *Darstellung*, der *Demonstration*, und nicht des Forschung ist. [...]
> Um was es sich mir bei meinem Forschungen handelt, ist die Zurückführung der complicirten Erscheinungen der Volkwirtschaft auf ihre wahren Ursachen, auf die constitutiven Factoren derselben und um die Erforschung der Gesetze nach welchen aus diesen letztern sich die complicirten Erscheinungen der Volkwirtschaft wieder aufbauen. Die Ergebnisse meiner Forschung können in mathem. Formeln gekleidet werden, mathem. Darstellungen deselben vermögen zu ihrer Demonstration beitragen: die mathem. Methode des Darstellung gehört indess keineswegs zu den Hauptaufgaben, welche ich mir gestellt habe. (Menger to Walras, 28/06/1883, p.768)[7]

Walras considered mathematics as a method of investigation and research (as did Jevons in the preface of his Theory of Political Economy (p. xviii)), while Menger only attributed to mathematics the role of representing the results of an investigation. There is here a crucial distinction between on the one hand the idea that mathematics is only a language, in line with the argument of Samuelson (1947) in favour of mathematical economics, and on the other hand that it can generate

[6] Karl Menger (1973) — Carl Menger's son — suggested a mathematical specification of the Austrian analysis, and argued that one of the main difference between Menger's analysis and the work of the other marginalists was that his father developed a concept of utility in terms of discrete variables — for which differentiability is not useful.

[7] However, I am not one of the true followers of the mathematical methods used in the treatment of our science. My thought is that the mathematical method is mainly a method of *exposition* and *demonstration* rather than of investigation. [...] The object of my research is to reduce complex economic phenomena to their true causes, and to seek out laws according to which these complex phenomena of political economy are repeated. The results of my research may be represented by mathematical formulae. Mathematical representations may help with the demonstrations: however, the mathematical method of representation is in no way the essential part of the task I have undertaken (translated by us and Gloria-Palermo (1999, p.33)).

independent methods of investigation, such as, according to Boumans (2004), the use of models in economics: this distinction is of a great importance for our investigation, since it already highlights the possible duality of the economic man, as the representation of a real individual or as a fictive and axiomatic individual, defined for investigation purposes.

Although the development of microeconomics seems to have been initiated by the work of engineers, Jevons, Menger and Walras had a more scientific purpose, and tried to establish more general economic principles than the study of specific situations such as the rail-road. We can for instance look at the preface to the first edition of the Theory of Political Economy, in which Jevons presents his conception of economics as 'a Calculus of Pleasure and Pain' (p.vii), or the discussion in Menger of the place of economics among natural sciences (pp.45-49), and also the explicit distinction Walras draws between science, art and ethics (2nd lesson).
In their attempt to produce a scientific knowledge of economy, they firstly needed to precisely define the scope of validity of their theories, i.e. the type of situations economics as a science is supposed to deal with. This definition was not necessary for engineers whose objective was the study of predefined subjects, such as the rail-road, but became essential in the enterprise of producing more general economic principles, without an *a priori* clear scope of validity.

### 2.2.2 The scope of economic analysis

An interesting similarity between Jevons, Menger and Walras is that they all had the same ultimate objective, the production of scientific laws of exchange: it is indeed the core phenomenon that must be studied within the context of a subjective theory of value. For instance, Jevons implicitly recognized that the aim of economic theory is the production of laws of exchange:

> The Theory of Economy thus treated presents a close analogy to the science of Statical Mechanics, and the Laws of Exchange are found to resemble the Laws of Equilibrium of a lever as determined by the principle of virtual velocities. (Jevons, 1871, p. vii)

Similarly, Menger — defending the legitimacy of investigating economic laws against the argument that human free will does not enable the existence of laws that could determine human behaviour — placed the explanation of exchange at the core of economics:

> Whether and under what conditions a thing is useful to me, whether and under what conditions it is a good, whether and under what conditions it is an economic good, whether and under what conditions it possesses value for me and how large the measure of this value is for me, whether and under what conditions an economic exchange of goods will take place between two economizing individuals, and the limits within which a price can be established if an exchange does occur — these and many other matters are fully as independent of my will as any law of chemistry is of the will of the practicing chemist. (Menger, 1871, p.48)

This objective was even clearer with Walras, who explicitly recognized that:

> For my part, I have done my utmost in the present half-volume to give a very thorough account of the mathematical theory of exchange (Walras, 1874, p.36)

The main difference between their works is that they did not adopt the same strategy in order to study the phenomenon of exchange: while Jevons and Menger tried to explain why the individuals are driven to exchange, Walras studied the exchange itself, and paid little attention to the underlying motives of the individuals.

#### 2.2.2.1 Jevons and the 'lowest rank of feelings'

Several years before the publication of the *Theory of Political Economy*, Jevons argued in a 'Brief Account of a General Mathematical Theory of Political Economy' that a theory of exchange can only be deduced from the fundamental determinants of individual behaviours:

> 2. A true theory of economy can only be attained by going back to the great springs of human action — the feelings of pleasure and pain. [...]
> 13. We now arrive at the theory of exchange, which is a deduction from the laws of utility. (Jevons, 1866, pp.282-285)

Jevons adopted a utilitarian perspective by considering that '[his] theory [...] is entirely based on a calculus of pleasure and pain; and the object of Economics is to maximise happiness by purchasing pleasure, as it were, at the lowest cost of pain' (Jevons, 1871, p.23), defining pleasure as 'any motive which attracts us to a certain course of conduct' and pain as 'any motive which deters us from that

conduct' (p.26). Such a definition implies that human action necessarily results from a calculus of pain and pleasure, but it also implies that we cannot homogenize and simply aggregate pleasures and pains, since 'a single higher pleasure will sometimes neutralise a vast extent and continuance of lower' (p.27). Therefore Jevons assumed the existence of a hierarchy of feelings, and specifically assigned to economics the study of the actions that result from 'the lowest rank of feeling' (p.27), i.e. the accumulation of wealth:

> Each labourer, in the absence of other motives, is supposed to devote his energy, to the accumulation of wealth. A higher calculus of moral right and wrong would be needed to show how he may best employ that wealth for the good of others as well as himself. But when that higher calculus gives no prohibition, we need the lower calculus to gain us the utmost good in matters of moral indifference. There is no rule of morals to forbid our making two blades of grass grow instead of one, if, by the wise expenditure of labour, we can do so. And we may certainly say, with Francis Bacon [about the riches], 'while philosophers are disputing whether virtue or pleasure be the proper aim of life, do you provide yourself with the instruments of either'. (Jevons, 1871, p.27)

Jevons clearly referred to an instrumental notion of reasoning, since he considered that economics should not discuss the ends and motives of the individuals — this is the work of moral philosophers — but only the instruments they use in order to achieve those ends. Jevons, quite similarly to Mill (1882, pp.1092-1093), isolated a specific economic motive in human behaviour, the accumulation of wealth: economics was therefore defined as the science of human actions undertaken in order to accumulate wealth.

#### 2.2.2.2 Menger and the satisfaction of human needs

Quite similarly to Jevons, Menger tried to explain the exchange from the behaviour of the individuals:

> A correct theory of price must instead be directed to showing how economizing men, in their endeavor to satisfy their needs as fully as possible, are led to give goods (that is, definite quantities of goods) for other goods. (Menger, 1871, pp.193-194)
> For economic theory is concerned, not with practical rules for economic

> activity, but with the conditions under which men engage in provident activity directed to the satisfaction of their needs. (Menger, 1871, p.48)

Menger's aim was indeed to understand the emergence and evolution of spontaneous and complex economic phenomena (Arena and Gloria-Palermo, 2001). His approach was then:

> to reduce the complex phenomena of human economic activity to the simplest elements that can still be subjected to accurate observation, to apply to these elements the measure corresponding to their nature, and constantly adhering to this measure, to investigate the manner in which more complex phenomena evolve from their elements according to definite principles. (Menger, 1871, pp.46-47)

However, unlike Jevons who specified a specific economic motive (and therefore assigned to economics the study of a relatively restricted range of human actions), Menger attributed to economics the role of explaining the efforts provided by the individuals in order to satisfy their needs, without any limitation on the set of needs economics is dealing with. Indeed, social phenomena 'must obviously have developed at some time from [their] simpler elements; a social phenomenon, at least in its most original form, must clearly have developed from individual factors' (Menger, 1883, p.149). It is therefore by studying how the individuals try to satisfy their needs that we will be able to explain *in fine* the emergence and evolution of institutions:

> If we summarize what has just been said we obtain the following propositions as the result of our investigation thus far: The principle that leads men to exchange is the same principle that guides them in their economic activity as a whole; it is the endeavor to ensure the fullest possible satisfaction of their needs. (Menger, 1871, p.80)

#### 2.2.2.3 Walras and the mathematical theory of exchange

Unlike Jevons and Menger who suggested grounding the analysis of exchange on theories of individual behaviour, Walras's purpose was the study of the mechanism of exchange itself. His main concern was indeed the study of the *social wealth*:

> By social wealth I mean all things, material or immaterial (it does not matter which in this context), that are scarce, that is to say, on the one

> hand, useful to us and, on the other hand, only available to us in limited quantity. [...]
> I say that things are useful whenever they can be put to any use at all; whenever they are seen to be capable of satisfying a want. [...] Furthermore, we need not concern ourselves with the morality or immorality of any desire which a useful thing answers or serves to satisfy.[...]
> I say that things are available to us only in a limited quantity whenever they do not exist in such quantities that each of us can find at hand enough completely to satisfy his desires. (Walras, 1874, p.65)

The determination of the value of the social wealth falls within the framework of science — i.e. the study of the truth — whereas the study of its industrial production is treated in the *Études d'économie politique appliquée* (Walras, 1898) and of its repartition in the *Études d'économie sociale* (Walras, 1896). The aim of pure economics is the study of the *exchange value* of the social wealth, which is determined by the market:

> Things that are valuable and exchangeable are also known as commodities. The market is a place where commodities are exchanged. Thus the phenomenon of value in exchange manifests itself in the market, and we must go to the market to study value in exchange. [...] In fact, the whole world may be looked upon as a vast general market made up of diverse special markets where social wealth is bought and sold. Our task is then to discover the laws to which these purchases and sales tend to conform automatically. To this end, we shall suppose that the market is perfectly competitive, just as in pure mechanics we suppose, to start with, that machines are perfectly frictionless. (Walras, 1874, §41)

We can therefore see that Walras represented the society as a vast market where the social wealth is exchanged, and assigned to pure economics the study of its exchange value. In addition, it clearly appears that Walras did not pay attention to the reasons of the exchange: he only observed that exchanges exist, and that they take place in the market. The focus is therefore quite different from Jevons and Menger, who studied a specific range of individual actions: while they were concerned about individual behaviours, Walras was concerned about prices (Peart, 1998, p.310). The main difference between their approaches is that Jevons and Menger — beyond their disagreements about the set of motives that could be defined as economic ones — focused on human *actions*, whereas Walras focused on

human *activities*. In the first case, the question of the definition of the scope of economics is a matter of human motives, and in the second case it is a matter of institutions.

We have shown in this section that the development of the marginalist thought with the independent and almost simultaneous publication of the works of Jevons, Menger and Walras gave a more scientific dimension to the microeconomics developed earlier in the 19th century, but that they defined different scopes of analysis for their respective theories, as well as different methodologies for a similar objective. The opposition between Jevons and Menger resulted from the difficulty of defining a clear set of motives to which an economic dimension can be linked, and Walras did not in fact adopt the same approach, and started his investigation from the simple observation that the phenomenon of exchange exists. In addition, whereas Jevons and Walras referred to the mathematical structure of economics developed by Cournot and the engineers, and also accepted the investigation dimension of mathematics — enabling the recourse to fictive mathematical entities in order to model the economy — Menger rejected the use of mathematics in economics.
We can already highlight a first difficulty within Jevons's approach: while Menger and Walras's use of mathematics are perfectly consistent with their ultimate objective, Jevons suggested studying individual behaviours and motives (calling for a quite realistic and psychological representation of the economic man) but argue in favour of mathematical models and justify them by the necessity of studying 'trading bodies' (Jevons, 1871, p.88), i.e. representative group of agents. The appeal to mathematical models and in particular differential calculus seems to be justified only when studying aggregate behaviours.

## 2.3 Individual behaviour in economic models

In this section, we show that the two approaches adopted by Jevons and Menger on the one hand, and by Walras on the other, generated two different representations of the 'economic man', one as part of a science of individual choice, and another as part of a science of social institutions.

### 2.3.1 Economics as a science of individual choice

We saw in the previous section that Jevons and Menger grounded their analysis of the exchange on individual behaviours: they therefore conceived economics as the

science of economic actions, i.e. a particular range of human actions, to which an economic dimension can be linked. This conception of economics as the study of economic actions can be traced back to Hume's *Treatise of Human Nature*: according to Demeulenaere (1996), Hume initiated an epistemological and methodological tradition in the social sciences that considers the analysis of social phenomena through a non normative approach (p.27). Hume indeed suggested a fundamental dichotomy between the 'passions', which determine the motives of any action, and reason, which is the ability of processing information about the reality. Hume then considered the passions as ultimate facts, and restricted the role of the reason to the representation of those passions through ideas:

> Since reason alone can never produce any action, or give rise to volition, I infer, that the same faculty is as incapable of preventing volition, or of disputing the preference with any passion or emotion. [...] Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them. [...]
> A passion is an original existence, or, if you will, modification of existence, and contains no representative quality, which renders it a copy of any other existence or modification. [...] 'Tis impossible, therefore, that this passion can be oppos'd by, or be contradictory to truth or reason; since this contradiction consists in the disagreement of ideas, consider'd as copies, with those objects, which they represent. (Hume, 1739, pp.266-267)

This representation of the working of human mind generated a specific epistemological tradition in social sciences — and more specifically in economics — in which the ends or tastes of the individuals are not discussed, but only the means they use in order to satisfy them (the primitive of an economic model is for instance systematically the *preferences* of the individuals). We have here the most fundamental characterisation of economic actions evoked in section 2.1, the association of an instrumental conception of rationality to this specific kind of actions: a necessary condition for a given action to be considered as an 'economic' one is that it results from an instrumental reasoning. We can however notice a difference between Hume's conception of the reason and the idea of an instrumental rationality. Hume indeed considered that an action is undertaken because there exists a passion that drives the individual undertaking this action, whereas an instrumental conception of rationality implies that an action is undertaken in order to satisfy a passion: the action results in the latter case from an *intention* of satisfying an objective, which is quite different from Hume's theory, according to which the

action results from the mere *manifestation* of the passion — manifestation which can be caused by the exercise of the reason and the discovery of some elements about the world (Sugden, 2006). This distinction is of a great importance for our inquiry since the notion of 'mistake', central in BWE, makes sense only with an instrumental conception of rationality, in which the individual can choose more or less appropriate means with respect to her objectives.
Hume's attempt was then to explain the emergence of social norms from the behaviour of individuals who seek to satisfy their interests: Jevons and Menger adopted a similar approach by analysing the exchange — and therefore the pursuit of a common advantage — from the behaviour of the individuals who seek either to accumulate wealth or to satisfy their needs. Similarly to Hume who lead a thorough investigation of the working of human mind, Jevons and Menger were concerned with the accuracy of their model of individual behaviour: Jaffé describes for instance the economic man in Menger's work as 'a bumbling, erring, ill-informed creature, plagued with uncertainty, forever hovering between alluring hopes and haunting fears, and congenitally incapable of making finely calibrated decisions in pursuit of her satisfactions' (Jaffé, 1976, p.521). Jevons was also aware of the complexity of the social phenomena[8], and explicitly endorsed Mill's concrete deductive method (Jevons, 1871, pp.16-17), considering that 'we may start from some obvious psychological law, as for instance, that a greater gain is preferred to a smaller one, and we may then reason downwards, and predict the phenomena which will be produced in society by such a law'. We can for instance find in the *Theory of Political Economy* some elements suggesting that Jevons was concerned with information issues and the uncertainty of future events (pp.35-36).

The different authors that continued the works of Jevons — Edgeworth (1881) and Pantaleoni (1889) for instance — and Menger — the Austrian School — then produced more elaborated theories of individual behaviours, but without the will of producing exclusively theories of exchange: unlike Jevons and Menger who suggested a theory of individual behaviour as a means to explain the exchange, some of their followers produced theories of individual behaviours as such. We can for instance see this shift in the work of Edgeworth, when he stated his 'first principle of Economics':

> The first principle of Economics is that every agent is actuated only by

[8]This is for instance the position of Peart, who argues that 'Jevons's view of human behavior is more complex than has been allowed, and has much in common with Menger's predisposition for process, uncertainty, mistakes, and the significance of time in decision making' (Peart, 1998, p.307).

> self-interest. The workings of this principle may be viewed under two aspects, according as the agent acts without, or with, the consent of others affected by his actions. In wide senses, the first species of action may be called war; the second, contract. (Edgeworth, 1881, pp.16-17)

While Jevons intended to study the exchange of goods, by assuming that the individuals are motivated by the accumulation of wealth, Edgeworth seemed to suggest studying any human action guided by the pursuit of self-interest: we can for instance notice that Edgeworth pointed out that the set of economic actions are divided into two categories, war and contract. This definition therefore recognizes apparently non economic activities — as we commonly understand it — as activities to which economic analysis can be applied, and for which neither exchange nor accumulation of wealth is involved.
We can find a similar evolution with Pantaleoni, who suggested defining economics as 'the laws of wealth systematically deduced from the hypothesis that men are actuated exclusively by the desire to realise the fullest satisfaction of their wants, with the least possible individual sacrifice' (Pantaleoni, 1889, p.3). There is here a will of providing theories of human behaviour, and not only laws of exchange.

This shift from a theory of exchange to a theory of human action is even more obvious in the pursuit of the work of Menger, and especially with the praxeology developed by von Mises:

> For a long time men failed to realize that the transition from the classical theory of value to the subjective theory of value was much more than the substitution of a more satisfactory theory of market exchange for a less satisfactory one. The general theory of choice and preference goes far beyond the horizon which encompassed the scope of economic problems as circumscribed by the economists from Cantillon, Hume, and Adam Smith down to John Stuart Mill. It is much more than merely a theory of the 'economic side' of human endeavors and of man's striving for commodities and an improvement in his material well-being. It is the science of every kind of human action. Choosing determines all human decisions. [...] The modern theory of value widens the scientific horizon and enlarges the field of economic studies. Out of the political economy of the classical school emerges the general theory of human action, praxeology. The economic or catallactic problems are embedded in a more general science, and can no longer be severed from this connection. No

> treatment of economic problems proper can avoid starting from acts of choice; economics becomes a part, although the hitherto best elaborated part, of a more universal science, praxeology. (von Mises, 1949, p.3)

We have therefore a first tradition in microeconomics in which we study the behaviour of the individuals, and not only the exchanges in which they are involved. Economics is defined as a science of individual choice, i.e. the science of human actions undertaken in order to satisfy some given ends. We can notice that it is mainly on the definition of the set of 'economic motives' that the economists from this first tradition disagree: Jevons only recognized the desire of accumulating wealth as an economic motive, Edgeworth and Pantaleoni extended the set of economic actions to the set of actions undertaken in order to satisfy one's interest, Menger was interested in the satisfaction of human needs in general, von Mises developed an apriorist methodology according to which any human action tautologically corresponds to the pursuit of an advantage[9], and therefore to the mobilisation of means in order to satisfy an end, and we can also mention the possibility of describing any human action as if it resulted from an instrumental reasoning and a deliberate calculus of maximisation of the advantage, in the line of the methodological principles suggested by Friedman (1953).

### 2.3.2 Economics as a science of social institutions

As mentioned above, Hume's analysis initiated a specific tradition in social sciences that considers the behaviour of the individuals through an instrumental view of reason — although Hume did not offer an instrumental theory of rationality — and that tries to explain social phenomena from the behaviour of the individuals. However, there exists another possible approach so as to study economic and social outcomes, which does not start from the behaviours of the individual, but from mechanisms at the scale of the society: one of the first authors having suggested such an approach is Adam Smith, through his analysis of the division of labour. Whereas Jevons and Menger can be situated in the continuation of the methodological tradition generated by Hume's analysis of the individual — although they differed in their

[9]This idea was already underlined by Aristotle, who stated that '[e]very art and every investigation, and similarly every action and pursuit, is considered to aim at some good' (Aristotle, 50BC, p.2), as well as Hobbes, according to whom 'of the voluntary acts of every man, the object is some good to himself' (Hobbes, 1651, pp.81-82). The existence of an action presupposed an intention in undertaking this action, and then the existence of a valuable reason from the point of view of the decision maker. The tautological dimension of this proposition led von Mises to consider it as an *a priori* truth.

characterisations of an economic dimension of human action — Walras is clearly in the continuation of Smith's analysis of the society, since he preferred to study the social wealth through the mechanism of the market and the exchange rather than the behaviour of the individuals. Walras's conception of economics — as the study of the exchange of the social wealth in the whole society, considered as a large market — is indeed quite similar to the one developed by Smith in the *Wealth of Nations*: Smith considered that the value is fundamentally determined by labour, which implies that the study of the wealth of the nations consists in the study of the repartition of the different activities through the division of labour. Smith therefore did not focus on a specific range of economic activities or human actions, but on the repartition of the labour — and therefore of the wealth — in the society. Walras did not intend to explain the origins of the division of labour, and considered it as a *natural fact*[10]. Furthermore, the notion of social wealth in Walras's work integrates the study of non material goods and activities within economics, quite similarly to Smith who considered those activities — such as philosophical or speculative functions (Smith, 1776, p.17) — as the product of the division of labour.

Contrary to Jevons and Menger who seemed to initiate a research program according to which economists should produce more and more complex theory of individual behaviour, Walras focused on the market equilibrium, and paid little attention to the behaviour of isolated individuals: the relevant scale of analysis is the society and the aggregation of individual behaviours. Walras therefore did not need to produce realistic or complex model of human behaviour: since the relevant phenomenon from the point of view of the economist is the aggregate behaviour, the individual elements that constitute it are of little importance. We can for instance refer to Walras's discussion about the continuity of the demand curve: Walras considered that its continuity is very unlikely at the individual level, but that it should be true for the aggregate demand (Walras, 1874, §52). If we refer to the discussion about the axiomatic characterisation of the economic man mentioned in the introduction, we can here notice that Walras did not try to explain the behaviour of an economic man from the behaviour of a real individual in the market, but from the aggregate demand, i.e. a property of the institution he is studying.
We can also notice a very poor description of individuals' motives in the work of

[10] Walras indeed considered that all economic relations are 'based on the natural fact that man's wants always surpass his own abilities to fulfil them' (Jolink, 2005, p.46): this implies that men tend to specialise in certain activities so as to share *in fine* the surplus of production with others in society. The division of labour is therefore a natural fact resulting from this natural propensity of men.

other economists whose object of study was market equilibrium, such as Cournot, who considered the act of exchange as a 'natural' and 'instinctive' act (Cournot, 1838, p.2), without further investigation. In addition, he also referred to the aggregative property of the demand function to justify its continuity (Cournot, 1838, pp.52-53). In the *Wealth of Nations*, Smith also provided an elementary description of the motives of the individuals engaged in exchange and argued that their study did not belong to his subject of investigation:

> This division of labour, from which so many advantages are derived, is not originally the effect of any human wisdom, which foresees and intends that general opulence to which it gives occasion. It is the necessary, though very slow and gradual, consequence of a certain propensity in human nature, which has in view no such extensive utility; the propensity to truck, barter, and exchange one thing for another. Whether this propensity be one of those original principles in human nature, of which no further account can be given, or whether, as seems more probable, it be the necessary consequence of the faculties of reason and speech, it belongs not to our present subject to inquire. It is common to all men, and to be found in no other race of animals, which seem to know neither this nor any other species of contracts. (Smith, 1776, pp.20-21)

We can therefore contrast two different approaches in economics: in the first one, economists try to explain why the individuals are brought to exchange, whereas in the second one, economists start their investigation from the observation that the individuals want to exchange, and then try to explain how the individuals actually exchange. This opposition has been for instance quite explicitly stated by Marshall:

> It is not true therefore that 'the Theory of Consumption is the scientific basis of economics'. For much that is of chief interest in the science of wants, is borrowed from the science of efforts and activities. These two supplement one another; either is incomplete without the other. But if either, more than the other, may claim to be the interpretor of the history of man, whether on the economic side or any other, it is the science of activities and not that of wants [...]
> From this it follows that such a discussion of demand as is possible at this stage of our work, must be confined to an elementary analysis of an almost purely formal kind. The higher study of consumption must come after, and not before, the main body of economics analysis; and, though

> it may have its beginning within the proper domain of economics, it cannot find its conclusion there, but must extend far beyond. (Marshall, 1890, pp.76-77)

We have here the explicit opposition we mentioned above, i.e. on the one hand economists who study the motives (or 'wants' for Marshall) of the individual, such as Jevons and Menger, and on the other economists who study human activities, and more broadly social institutions[11]. The characterisation of the economic man is quite different within this second approach: the economists do not need an elaborated theory of individual behaviour, and will therefore refer to a representative agent of the group of agents they are studying. It seems that it is Marshall — who defended a bit paradoxically the view that economists should 'deal with man as he is: not with an abstract or "economic" man; but a man of flesh and blood' (p.22) — who firstly used the notion of a representative agent:

> So far we have looked at the demand of a single individual. And in the particular case of such a thing as tea, the demand of a single person is fairly representative of the general demand of the whole market [...].
> There are many classes of things the need for which on the part of any individual is inconstant, fitful, and irregular. There can be no list of individual demand prices for wedding-cakes, or the service of an expert surgeon. But the economist has little concern with particular incidents in the lives of individuals. He studies rather 'the course of action that may be expected under certain conditions from the members of an industrial group,' in so far as the motives of that action are measurable by a money price; and in these broad results the variety and the fickleness of individual action are merged in the comparatively regular aggregate of the action of many. (Marshall, 1890, pp.82-83)

> We shall have to analyse carefully the normal cost of producing a commodity, relatively to a given aggregate volume of production; and for this purpose we shall have to study the expenses of a representative producer for that aggregate volume (Marshall, 1890, p.264)

[11] It should be noticed that there also existed some tensions in this second tradition concerning the scope of economic analysis: whereas Walras intended to study the social wealth in its totality, Marshall explicitly restricted economics to the study of a precise range of human activities, the ordinary business of life: 'Political Economy or Economics is a study of mankind in the ordinary business of life; it examines that part of individual and social action which is most closely connected with the attainment and with the use of the material requisites of wellbeing' (Marshall, 1890, p.1).

Unlike the research programs of Jevons and Menger, the priority is not here the elaboration of a theory of individual behaviour, but the study of the interactions of large groups of individuals. In this second situation, economics is conceived as a science of institutions, and not of individual choice any more: the phenomenon the economists study is not the determinants of individuals' behaviours, i.e. the reasons why the individuals act, but the interaction between large groups of individuals. From a methodological point of view, it is therefore probably not relevant to produce complex models of individual behaviour which will not be easily tractable for the study of aggregate behaviours. Economists should therefore define 'average' agents, presenting only the key characteristics of the group of economic agents they represent: the economic man is therefore not an idealisation of a real individual — which would integrate the main features of a real individual — but an idealisation of a group of individuals. It is then possible to attribute to this representative individual the properties observed at a higher scale, such as continuous demand curves.

The work of Schelling (1978) is a good illustration of this tradition: his attempt is indeed to explain observed social outcomes from the behaviour of isolated individuals, who are explicitly designed for the explanation of a specific institution (such as racial segregation). Schelling does not try to explain social phenomena from a unique model of the individual, but designs a specific model of the individual such that it explains a given social phenomenon: the economic agent is conditioned by the institution the economist wants to study. This approach directly derives from Friedman (1953) 'as-if' methodology: in his famous explanation of the density of leaves around a tree as the result of an explicit calculus of maximisation by the leaf itself (who can decide its own position so as to maximise the amount of sunlight it receives), the theory can be validated because it is consistent with the resulting outcome. The model of individual behaviour is therefore defined so as to explain *in fine* the social outcome. Friedman and Savage (1948) made earlier a similar point by considering that assuming that the shots of an expert billiard player can be described as the result of a complex and explicit optimisation problem is a reasonable model: it is justified *a posteriori*, since, although billiard players are not able to make the required calculus, the reason why they are expert is that they found a way to reach the exact same outcome. The model of the representative billiard player is therefore defined from the outcome of the decision: there is in particular no reason that this representative agent may offer insights about how *real* people behave.

In this section, we showed that the marginalist revolution produced two differ-

ent models of the economic man. The first one, initiated by Jevons and Menger, intends to describe the behaviour of actual individuals and is concerned with the determinants of individual choice. The individuals are facing numerous information and cognitive issues, and are supposed to be trying to achieve some given ends. The second model, initiated by Walras, uses a notion of representative individual in order to get a simpler representation of the society, as the interaction between ideal and fictive individuals. The object of economics is then the study of social phenomena such as the formation of market prices.

## 2.4 Pareto and the *Homo economicus*

We can distinguish between two different conceptions of the economic man, according to the objective of the economist: either an institution-based model, that models a representative individual from the characteristics of the institution the economist studies, or an individual-based model, concerned with the realism of its representation of the individual. We suggest now that Pareto homogenized these two conceptions with the figure of the *Homo economicus*.

### 2.4.1 Pareto's science of logical actions

By opposition to the works of the early neoclassical authors such as Jevons, Edgeworth and Pantaleoni, Pareto wished to eliminate all psychological interpretations within economics, in order to establish a theory based on principles of rational choice. Pareto considered that '[c]learly psychology is fundamental to political economy and all the social sciences in general' (Pareto, 1909, chap.2; §1), and therefore that — in the future — all social phenomena may be deduced from principles of psychology (which can also be deduced from principles of biology, and then from principles of chemistry and physics). Our current lack of knowledge however forces social scientists to adopt an alternative road: Pareto therefore suggested a classification of human actions in order to simplify the scientific analysis of human society. He grounded his classification on the existence or not of a logical connection between the action and the objective of the individual, and if this logical connection is objective or only subjective. Pareto then defined logical actions as the class of human actions such that (1) the action is undertaken in order to satisfy a given objective and (2) this action is objectively appropriate towards the objective of the individual. Therefore logical actions 'logically conjoin means to ends not only from the standpoint of the subject performing them, but from the standpoint of other persons

who have a more extensive knowledge' (Pareto, 1916, §150). Pareto then seemed to define economic actions as the range of logical actions for which the objective of the individual is the satisfaction of her tastes through the purchase of goods:

> We will study the many logical, repeated actions which men perform in order procure the things which satisfy their tastes. [...] Moreover, we will simplify the problem still more by assuming that the subjective fact conforms perfectly to the objective fact. This can be done because we will consider only repeated actions to be a basis for claiming that there is a logical connection uniting such actions. A man who buys a certain food for the first time may buy more of it than is necessary to satisfy his tastes, price taken into account. But in a second purchase he will correct his error, in part at least, and thus, little by little, will end up by procuring exactly what he needs. We will examine this action at the time when he has reached this state. [...]
> 2. Thus by considering only one part of man's actions and, in addition, by assigning certain characteristics to them, we have simplified the problem enormously. The study of these actions makes up the subject of political economy. (Pareto, 1909, chap. 3, §1)

Nevertheless, he also recognized the possibility of producing economic theories without this specific restriction of motives:

> [S]ince it is customary to assume that man will be guided in his choice exclusively by consideration of his own advantage, of his self-interest, we say that this class [of economic theories[12]] is made up of theories of egotism. But it could be made up of theories of altruism (if the meaning of that term could be defined rigorously), or, in general, of theories which rest on any rule which man follows in comparing his sensations. It is not an essential characteristic of this class of theories that a man choosing between two sensations chooses the most agreeable; he could choose a different one, following a rule which could be fixed arbitrarily. (Pareto, 1909, chap.3, §11)

Economics seems therefore to be the science of logical actions in general, although Pareto previously restricted the set of economic actions to the pursuit of

[12]Pareto made a distinction between theories on interpersonal and intrapersonal comparisons of utility, and refers in this extract to the second class of theories.

a specific motive, the satisfaction of one's tastes through the purchase of goods. In the continuation of Mill's epistemological principles, Pareto suggested then the concept of the *Homo economicus* as the 'dimension' of the individual that deals with economic actions, by analogy with the study of a concrete body, which can be seen under different perspectives (chemical, geometrical...). Therefore in reality man is an aggregate of all these different dimensions (the *Homo economicus*, the *Homo religiosus*, the *Homo ethicus*...) and apart from a few very specific contexts, the behaviour of an actual man is most of the time different from the behaviour of the *Homo economicus*:

> For certain concrete phenomena the economic side matters more than all the others. In such a case, one can, without serious errors, restrict himself to the results of economic science alone. There are other concrete phenomena in which the economic side is insignificant, and there it would be absurd to restrict oneself to the results of economic science alone. Quite the contrary, they should be disregarded. There are intermediate phenomena between those two types; and economic science will reveal a more or less important aspect of them. In all cases, it is a question of degree, more or less. (Pareto, 1909, chap.1, §27)

Pareto placed himself in the continuation of the first tradition initiated by Jevons and Menger, since he tried to produce a theory of human action, and defined the scope of validity of economics by a specific set of motives[13]. However, Pareto's objective was the study of social phenomena, such as the circulation of elites, and he had then to characterize this *Homo economicus* consistently with this objective. He claimed for instance that a psychological description of the individual is not acceptable: the *Homo economicus* being an abstract entity of pure economics, it will obviously not fully describe the behaviour of a real individual — which is conditioned by psychological and sociological factors — but enables economists to roughly describe the mutual dependence of economic phenomena[14] (Pareto, 1916, §36). Since Pareto's purpose is the study of social institutions — in the *Manual*

[13] Pareto however seems to oscillate between Jevons and Menger's position, since he argued on the one hand that economic actions are performed towards the satisfaction of one's tastes and on the other hand that the only relevant phenomenon is the satisfaction of one's ends, whatever they are.

[14] According to Pareto, if we accept the idea that economics should integrate psychology in order to be more empirically accurate, then there would be no reason to describe economic phenomena only thanks to psychology, since they also depend on geography and more generally on all natural sciences. Pure economics does not have the ambition of explaining all social phenomena, but only a certain part of them, for which the individuals performs logical actions. Since pure economics only

*of Political Economy*, he indeed explicitly recognized that the main purpose of his work is the study of economic equilibrium (Pareto, 1909, chap.3, §14)[15] — he in fact did not need a psychology-based model of the individual: the man he suggested to split in different *Homines* is not a real individual, but a representative one. Following a logico-experimental method, he argued that the properties of this representative individual should not be characterized *a priori*, but from the observed social phenomena (Pareto, 1916, §55, §63). In particular, the *Homo economicus* should be described from the properties of the social institution in which the representative individual tries to satisfy her tastes through the purchase of goods, i.e. from the properties of repeated markets.

Similarly to Jevons and Menger — who defined the economic man as the idealisation of a real individual —, Pareto identified a specific set of actions for which the economic man offers a good approximation of the behaviour of a real individual, but similarly to Marshall — who defined the economic man as the idealisation of a group of individuals — he defined the economic man as a representative agent. We have therefore the homogenisation of the two distinct models of the economic man: the *Homo economicus* is indeed simultaneously a representative individual of the market and a real individual who learned how to satisfy her tastes. Although Pareto tried to produce a theory of individual behaviour grounded on principles of rational choice rather than psychology, his main objective was the analysis of social phenomena: he did not actually need to produce such a model of individual behaviour, and could simply have defined a model of a representative agent characterized by the properties of repeated markets.

### 2.4.2 *Homo economicus* and the economic way of looking at behaviour

The ambiguity of the definition of the neoclassical economic man can be traced back to the paretian *Homo economicus*. Pareto indeed defined a *Homo economicus* for the study of institutions, but tried to justify his model from properties of individual choice, such as learning and the discovery of one's own tastes. This ambiguity

needs a theory of logical actions and not of human actions in general, psychology can be removed from economics.

[15] We can also notice that Pareto emphasised that economists are studying the behaviour of large number of individuals, and therefore that they should focus on average behaviours, rather than individual ones (see for instance Pareto (1909, chap.3, §65-66, §87) and Boianovsky (2013, pp. 109-111).

led economists to assume that real individuals were characterised by underlying coherent preferences: an individual progressively discovers her true preferences, and economic theory is predicting the behaviour of this individual when she has reached this state. This idea is indeed explicitly stated by Pareto in his definition of the scope of economics: the repetition of the action enables the individual to learn how to satisfy her tastes, which are supposed to be well-ordered.
The observation of preference incoherences in laboratory experiments therefore only means that people have not discovered their true preferences yet. This perspective, grounded on the duality between on the one hand the rational *Homo economicus* with coherent preferences (the *inner rational agent*) and the actual individual with her incoherent preferences, has been explicitly endorsed by the proponents of the *discovered preferences hypothesis* (Smith, 1989, 1994, Plott, 1996): the repetition of the game enables the individual to get more experience and to discover what she truly wants, and *in fine* to perform logical actions and act like the *Homo economicus*. Since this rationalisation process is not related to a specific institution but to the mere phenomenon of learning, it seems possible to explain any human action thanks to rational choice theory, as long as the individuals are able to learn from their past choices. Economists then removed all psychological considerations from their theories, and, in the second half of the 20th century, started to apply the model of the *Homo economicus* to explain human actions undertaken outside repeated markets. This was for instance the intention of Becker, who suggested interpreting the *Homo economicus* as the economic way of looking at human behaviour (Becker, 1993).

This conception of the individual — as an *Homo economicus* who progressively frees herself from her incoherences — implies that the predictive failures of rational choice theory can be explained by a transient state of 'non rationality' of the individual. It is therefore because the individual has not discovered her true preferences yet that the theory does not predict well her choices, and not because the theory is false. This is precisely the argument used by Binmore in order to question the relevance of behavioural findings: the discovery of anomalies in human behaviour in the laboratory does not question the validity of the theory, since the individuals are not in real economic conditions. This raises the issue of the definition of the scope of validity of rational choice theory as a theory of individual behaviour. We can indeed notice that the set of logical actions is ill-defined: it is not defined by objective criteria such as the purchase of a good, but — if we refer to Binmore's conditions — by criteria such as *adequate* incentives or a *sufficient* time for trial-and-error adjustments. In particular, even though it is possible to create experimental settings

in which the individuals become individually rational, it is not certain at all that those conditions are verified in real markets.

### 2.4.3 Rational choice and preference shaping

The representation of the individual as an economic man who progressively becomes rational is grounded on the assumption that learning enables the individual to discover her true, coherent preferences. Plott (1996) then argues that repeated markets provide a suitable framework for the discovery of those true preferences. However, the fact that people tend to become individually rational in specific market experiments characterized by learning is not sufficient to deduce that the rationality of the individual is only due to the learning effect: there maybe exists a specific process related to repeated markets experiments that *shapes* individual preferences. Loomes et al. (2003) designed for instance an experimental setting in order to show that repeated markets can shape the preferences of the individual: even though there can exist a learning effect that helps people to correct their past mistakes, they show that systematic shaping effects occur (Loomes et al., 2003, p.C166). Imagine for instance the following situation:

> **Isidore's intransitive preferences**: consider an individual with intransitive preferences, Isidore, who prefers the good A over the good B, B over C and C over A (he is ready to buy one unit of B (resp. C, A) against €1 plus one unit of A (B, C)). Since a market gives him the opportunity to exchange his initial endowment so as to obtain goods that he prefers, we can see that Isidore is likely to be victim of a money pump. After a few transactions, he becomes aware of this inconsistency (and of the risk associated to it), and decides not to sell any more the good C if C is in his possession.

By repeatedly exchanging on the market, and therefore by being victim of a money pump, Isidore's preferences changed: it is therefore the *nature* of a specific institution (repeated markets), that caused Isidore's change of preferences, and not the discovery of some latent coherent preferences. Isidore could have for instance decided not to sell any more A or B instead of C: the fact that he decided not to sell C any more is probably path-dependent. He may for instance have started his series of exchange by buying 1 unit of B against €1 and 1 unit of A, and then sold this unit of B plus €1 so as to get one unit of C: Isidore then became aware that buying one unit of A for one unit of C plus €1 would reveal an incoherence (since he would be back to his initial endowment minus €3). He can then decide not to sell

the good C so as to regain coherent preferences, and not be fooled by a money pump.

In the specific context of repeated markets experiments, we could either interpret the process of rationalisation of Isidore's preferences as the result of learning or shaping. There is however no reason to believe that another social institution will shape coherent preferences — whereas learning necessarily leads to the discovery of coherent preferences. Unless economists can show that the rationalisation of individual preferences in repeated markets experiments is only due to learning and not to preference shaping — and therefore that learning enables any individual to discover her true preferences in any social situation — it is quite dubious to refer to the model of the *Homo economicus* to explain human actions outside markets: there is indeed a possibility that the preferences are shaped by the institution itself. This would mean that such 'true' coherent preferences do not exist, and that one's revealed preferences are not the result of a combination of true preferences and mistakes. Consider as an illustration the Christmas cards institution (Schelling, 1978):

> **Christmas cards**: individuals send Christmas cards to their relatives, mainly as the result of an interactive process 'greatly affected by custom and by expectations of what others expect and what other may send, by cards received (and not received) last year and already received this year' (p.31). One of the main features of this institution is that people 'feel obliged to send cards to people from whom they expect to receive them, often knowing that they will receive them only because the senders expect to receive cards in return' (pp.31-32). This institution leads to suboptimal results, and '[t]here is no mechanism that would induce people to stop sending cards merely because everybody, like everybody else, deplored the system and wished it would disappear' (pp.32-33).

We have here a social institution satisfying the property of learning (the individuals can learn from the past years who is likely to send them a card this year), but grounded on incoherent preferences. We can indeed notice that there exist a significant number of individuals who respects the tradition by sending cards, but who would also prefer not to send nor receive any card. The institution has therefore implemented non transitive preferences, since — after the repetition of the interaction each year — the individuals still prefer to respect a tradition whose disappearance is preferred to its preservation. This means that individuals' preferences are probably greatly influenced by the institutional framework within

which they interact, and therefore that the phenomenon of rationalisation in repeated markets is maybe more a property of the institution itself of repeated markets than of learning and the discovery of underlying coherent preferences. In particular, it may be not legitimate to extend the model of a rational *Homo economicus* to study phenomena out of the market, since it is not certain that another institution than repeated markets could shape coherent preferences.

Since economists imputed the rationalisation of individual preferences observed in repeated markets to learning, they started to study individual behaviours out of the market, and used an inappropriate model of individual behaviour. The model of a rational *Homo economicus* — initially thought as a model of a representative individual — was then not empirically relevant, and it became necessary to change the economists' tools: it was therefore quite logical that some economists started to reuse the methodology of the earliest neoclassical authors in order to produce a model of the individual designed for the study of human actions — and not only of the specific institution of repeated markets. This is for instance the position of Bruni and Sugden (2007), who argue that the current opposition between behavioural and neoclassical economics can be interpreted as the re-emergence of the debate between Pareto and the economists from the first tradition — Jevons, Edgeworth and Pantaleoni — about the validity of explaining individual behaviours thanks to axiomatic principles of rational choice rather than psychology based theories. Our claim is that referring to principles of rational choice can be legitimate if we want to characterize a representative individual — under the condition that the actual social outcome can be modelled as the result of the interaction of individually rational agents, such as repeated markets or racial segregation — but is much more delicate if we want to model the behaviour of a real individual, since the scope of validity of the theory is ill-defined and the preferences can be shaped by social institutions. It is indeed worth noticing that the discovered preferences hypothesis has no firmly established psychological foundations. Plott's justification of this hypothesis is indeed that it provides a convenient primitive for economic models, and not that it is psychologically relevant:

> Psychologists tends to distinguish their work from what they call a 'philosophy of articulated values', as opposed to a 'philosophy of basic values', which psychologists tend to embrace. The former, sometimes attributed to economists by psychologists, would hold that people have well-formed preferences or values. Choices are then made by reference to these values, which themselves are stable. By contrast, psychologists

> see themselves as operating under a philosophy of basic values from which preferences might be viewed as 'constructed'. The construction depends upon the mode in which a response is called. Task and context are thought to influence the construction and, as a result, preferences are thought to be labile if, indeed, they can be said to exist at all. Of course, if no preferences exist, then there is no foundation for a theory of optimisation and no foundation for a theory of strategic behaviour and game theory. The idea of constructed preferences would seem to leave very little room for economics and seems to be substantially contradicted by the existence of economic models that are so powerful in applications. (Plott, 1996, p.227)

The reason why economists are committed to a notion of underlying coherent preferences (that may be progressively discovered by the individual) is therefore not because it provides an empirically relevant model of individual behaviour, but because it is a convenient model for economic analysis.

## 2.5 Conclusion

The main criticisms addressed towards behavioural economics concern its 'non-economic' approach. We however showed that the two arguments used against behavioural economics are grounded on two quite different conceptions of economics: one as a science of individual choice, and the other as a science of social institutions. We showed that this duality resulted from the development of modern microeconomics, Menger and Jevons claiming that economics should study human actions, whereas Walras and later Marshall argued that economics should study human institutions. Both approaches were then developed in parallel, until their homogenisation with Pareto and the *Homo economicus*: Pareto indeed intended to study social phenomena, and simultaneously referred to arguments of both approaches so as to characterize his *Homo economicus*. The shift from a model of the economic man grounded on psychology to a model grounded on principles of rational choice enabled economists to widen their field of investigation: since learning seemed to be the phenomenon that enables people to become rational by discovering their underlying coherent preferences, it was possible to apply this model to the study of any social institution. The proof of the existence of logical actions, for which the individuals have discovered their true preferences, is the objective of the proponents of the discovered preferences hypothesis: however, although it seems possible to create experimental settings in which the subjects

progressively becomes rational, this objective is pointless, since the assumption of rationality of the representative individual should not be deduced from the behaviour of actual individuals, but from the characteristics of the institution the economist is studying. Furthermore, the rationalisation of individual preferences may also be the result of preference shaping and not of learning.
Since the figure of the *Homo economicus* was designed for the study of economic equilibrium but also offered a tractable model of individual behaviour, economists started to study human actions which were not logical with an inadequate model. This extension of the scope of economics from the study of logical actions to the study of human actions in general implied a need for new economic tools, such that economists could model real individuals and not only the representative individual of repeated markets: the development of behavioural economics — suggesting the introduction of psychology based models of behaviour — was then a logical consequence of this evolution.

Against Levine's claim that behavioural economics is not studying the right range of phenomena, we argue that there do not exist any specific reason to restrict economics either to the study of individual choices or to the study of social institutions: this duality is indeed fundamental in economic analysis, since the first tradition suggests studying the determinants of individual behaviour — i.e. individual preferences — and the second tradition studies the interaction of large groups of individuals — for given preferences. Economists should therefore refer to the model of human behaviour which is the most adapted to their purpose. In particular, if economists are studying individual behaviours in settings for which they are not necessarily rational, then it seems necessary to provide psychology based theories of individual choice. The assumption that the individuals tend to become rational does not imply that rational choice theory is sufficient to model human behaviour: it only means that rational choice theory is relevant as a theory of individual behaviour in a restricted range of situations — much more restricted than the range of situations with which economists are dealing.
The issue that economists must solve is then the trade-off between the tractability and the empirical validity of their model of individual preferences. Economists concerned with individual behaviours — such as food behaviour or marriage — should therefore refer to complex but realistic models of human behaviour, whereas economists concerned with social institutions — such as the phenomenon of obesity or interracial marriages — should privilege a model of a representative individual, and deduce from the properties of the social phenomenon the preferences of this representative individual. This implies in particular that if we study a specific phe-

nomenon for which the individuals obviously present systematic decision flaws, or that can shape incoherent preferences, the representative individual will not necessarily be defined by coherent preferences. We can for instance think of the prospect theory (Kahneman and Tversky, 1979) as a good illustration of a model combining the tractable structure of rational choice theory and incoherent preferences: Kahneman and Tversky identified within the group of subjects of their experiments a common psychological characteristic, loss aversion; they then used the simple mathematical structure of rational choice theory in order to introduce their experimental findings, and produced *in fine* a relatively simple model of individual behaviour, with a more robust model of preferences. This can therefore constitute an interesting model of representative individual if we study a social phenomenon that could result from the interaction of isolated individuals characterized by loss aversion, or if there is a specific mechanism in the institution that implements loss aversion in the preferences of the individual. However, modelling a social outcome by the interaction of loss averse representative agents does not say anything about the preferences of the real individuals.

Chapter 3

# The Paretian Foundations of Behavioural Welfare Economics*

Contents




**Abstract:** Behavioural welfare economics aims at designing public policies helping boundedly rational individuals to satisfy their own preferences. It is assumed that (i) individuals are defined by true preferences on which they would act were they able to reason correctly, (ii) the satisfaction of those preferences constitutes the normative criterion, and (iii) it is possible to elicit those preferences from the social planner standpoint. We argue that this model was implicit in neoclassical economics from Pareto on, and highlight the conceptual difficulties of those three claims. We claim that those difficulties are due to the project of behavioural economists of integrating psychology into economic models, rather than grounding economic models on psychology. We illustrate those difficulties by discussing the Oscar-case.



*The section 3.6 of this chapter has been independently accepted for publication in the *Review of Philosophy and Psychology*. I thank Adrien Barton, Till Grüne-Yanoff and three anonymous referees for their comments and suggestions.

## 3.1 Introduction

As discussed above, the most common solution given by behavioural economists to the reconciliation problem is to assume that individuals are characterised by underlying 'true' preferences (like the ones we found in neoclassical economics) — the preferences that would be revealed by the choices of the individuals, were they able to reason correctly — and to treat deviations from the satisfaction of those preferences as mistakes (e.g. Köszegi and Rabin (2007, 2008), Bernheim and Rangel (2007, 2009), Rubinstein and Salant (2012)). The satisfaction of the underlying preferences is then taken as the normative criterion (Sunstein and Thaler, 2003, Thaler and Sunstein, 2008)). In particular, since the individuals are likely to make mistakes — by taking into account irrelevant features of the choice environment when comparing different alternatives — the social planner is legitimated to implement measures such that the individuals will be able to satisfy *in fine* their true preferences (Conly, 2013). We however suggested in chapter 2 that nothing justifies *a priori* that the notion of true preferences could be relevant outside repeated markets, in which it is possible to define *objectively* the true preferences of the individual (on an homogeneous market, the objective of any consumer is to buy the product at the lowest price). The aim of this chapter is to highlight that BWE adopts a methodology similar to the one developed by Pareto, but that — unlike Pareto — behavioural welfare economists intend to study non-market outcomes while keeping the notion of true preferences.

We argue that this approach does not solve the issues raised by behavioural findings, since behavioural welfare economists are simply trying to integrate psychological factors into economic models, without questioning the validity of the underlying economic model (and in particular the existence of true preferences). We firstly present the three main claims underlying behavioural welfare economics, i.e. (i) people are defined by true preferences, (ii) the satisfaction of those true preferences is the normative criterion, and (iii) it is possible to objectively elicit the true preferences of the individuals (section 3.2). We then suggest that this conception of the individual was already implicit in neoclassical economics from Pareto on, and successively question the validity of those three claims: we show that the existence of true preferences can be postulated if and only if we accept Pareto's reductionist account of reasoning (section 3.3), that the normativity of the satisfaction of one's true preferences is not a consequence of Pareto's theory and is not properly justified (section 3.4), and that economists do not have at their disposal an impartial mechanism that would allow them to elicit the true preferences of the individuals (section 3.5). We illustrate the difficulties of BWE by focusing on the Oscar-case

(section 3.6). We conclude by stressing that reconciling normative and behavioural economics requires grounding economic models on psychology instead of integrating psychology into economic models (section 3.7).

## 3.2 Preferences and mistakes

As an illustration of the issues for welfare economics raised by behavioural findings, imagine the following situation:

> **Luke's investment choice**: Luke intends to invest €100, and seeks guidance from his bank advisor Claire. Suppose firstly that Luke is given the choice between (A) investing his initial endowment in a safe asset, giving him a sure gain of €20, or (B) investing in a risky asset, giving him a gain of €60 with probability 1/3 and nothing with probability 2/3. Suppose now that, instead of having to choose between (A) and (B), Claire can offer €60 to Luke in a risky asset if he leaves his €100 in a non-remunerated account. Luke's choice is then between (C) reselling the asset at one third of its current value (i.e. losing for sure two thirds of its value), or (D) keeping the asset which is likely to be devaluated with probability 2/3 (his value is null in this case, while it still pays €60 with probability 1/3).

Suppose that Luke chooses (A) in the first case and (D) in the second (in line with experimental results on loss aversion, Tversky and Kahneman (1981)): since (A) and (C) (as well as (B) and (D)) are identical in terms of payoff distribution (which constitutes the only relevant element of the choice problem from a consequentialist perspective), those choices reveal that Luke's preferences changed when the framing of the problem changed. The common interpretation of this preference reversal is that — unlike a standard neoclassical agent — Luke's choices reveal a loss aversion. Luke is indeed risk averse when facing lotteries involving gains, but risk-seeking when facing lotteries with losses (Kahneman and Tversky, 1979): in the first situation, he prefers the sure gain to the risky one, while in the second situation — when given an initial endowment of €60 — Luke prefers the risky option (he is indeed comparing a sure loss of €40 with a probability of 1/3 of avoiding a €60 loss). The difficulty for welfare economics is then to know in which case preference satisfaction matters, since satisfying Luke's preferences in both situations leads to different choices — although the relevant features for the evaluation of the alternatives remain identical. The common interpretation of this kind of situation is that Luke truly prefers one option over the other (either (A)

over (B) and (C) over (D), or (B) over (A) and (D) over (C)), but that his actual choice is influenced by the framing of the decision problem.
Consider now the choice problem faced by Claire. Suppose that Claire noticed that the way she frames the investment choice has a significant influence on the final decision of her clients. She therefore knows that framing the problem as a choice between (A) and (B) will induce Luke to choose the safe asset, while he will probably choose the risky one if he must choose between (C) and (D). Suppose also that Claire is benevolent, i.e. she chooses the choice architecture of the problem such that Luke will be better off, from his own point of view. Claire's objective is therefore not to impose to Luke what she thinks is good for him, but to help him to get what he would have chosen if he was able to choose correctly, i.e. if he was not influenced by irrelevant features of the choice architecture. Suppose for instance that, thanks to past interviews, Claire can reasonably assume that Luke is actually a risk-lover, and that he would choose the risky option if he were not influenced by framing effects. Although she may believe that it would be better for Luke to take more cautious decisions (for the simple reason that she is risk averse herself), her benevolence implies that she will respect Luke's underlying preferences. She will therefore choose to frame the decision problem as a choice between (C) and (D), since it is more likely that Luke takes *in fine* the risky asset.

If we accept the assumption that Luke truly prefers one asset over the other — independently of the way Claire framed the choice problem — and that his actual choices may be influenced by elements he would consider to be irrelevant if he were aware of them (e.g. details of the choice architecture), then an intuitive normative criterion would be the satisfaction of Luke's true preferences. The core argument of BWE is that the individuals have true preferences they wish to act on, but that most of their decisions are influenced by irrelevant cues, such as the desire of avoiding losses, or more generally the greater salience of an alternative (in Sunstein and Thaler's example of the cafeteria, the individuals tend to choose more prominently-displayed items, independently of their nature). This dual model of the agent distinguishes between the true preferences of the individual on the one hand (whose satisfaction determines her level of well-being, which seems a normatively desirable objective) and her revealed preferences on the other hand (whose satisfaction determines her choice).
The individuals intend to make their life go as best as possible, but are not able to take the adequate decisions due to psychological biases. Köszegi and Rabin (2007, 2008) suggest for instance understanding individual behaviour as a combination of preference satisfaction and mistakes, and then argue that 'preferences often can be

*revealed* by behavior, even when they are not *implemented* by behavior' (2007, p.479, emphasis in original): it then falls to welfare economists to distinguish between the underlying coherent preferences of the individual and her mistakes. The individual is therefore a rational agent who tries to satisfy her coherent preferences, but who can fail to satisfy them due to psychological biases. The agent is maximizing a function corresponding to her true preferences distorted by a set of parameters representing her psychological biases, and is therefore maximizing a 'wrong' utility function (in the sense that what drives her behaviour is not the satisfaction of her true preferences, but the maximisation of a distorted utility function). In a discussion in the *Journal of Economic Literature* with Harstad and Selten (2013) and Crawford (2013), Rabin (2013) explicitly defends this 'optimisation approach' (in contrast to bounded rationality models, see Harstad and Selten (2013) for a review). His objective is to:

> making (imperfect and incremental) improvements over previous economic theory by incorporating greater realism while attempting to maintain the breadth of application, the precision of predictions, and the insights of the neoclassical theory (Rabin, 2013, p.528)

Behavioural economists should therefore consider the neoclassical model of behaviour as a benchmark, and then progressively integrate into individual utility functions different parameters (that can be interpreted as expressing specific psychological inclinations of the decision maker) so as to improve the predictive power of the theory. The primitive of the model therefore remains the standard neoclassical framework with coherent preferences.

So as to provide a viable answer to the reconciliation problem, behavioural welfare economists therefore states three more or less explicit hypotheses — which are actually standard hypotheses in conventional welfare economics:

(i) each individual is fundamentally defined by true preferences, i.e. a unique consistent and context-independent preference ordering which would determine her choices, were she able to reason correctly;

(ii) the satisfaction of those true preferences is the normative criterion;

(iii) it is possible to identify this preference relation from the choice architect's standpoint.

Hypothesis (i) constitutes the primitive of BWE and is therefore essential for the different political agendas relying on BWE, such as Sunstein and Thaler's libertarian paternalism. If hypothesis (i) is true but not hypothesis (ii), then nudging the individuals such that they satisfy *in fine* their true preferences cannot be considered as normatively desirable (and therefore LP's paternalistic claim would be ethically problematic). Lastly, if (i) and (ii) are true but not (iii), then it means that, although it would be desirable to nudge the individuals such that they satisfy their true preferences, the choice architect cannot elicit with certainty what those true preferences are. Therefore, if at least one of these assumptions is not verified, BWE's project of identifying individuals' underlying true preferences (so as to help the individuals to satisfy those preferences) is a dead-end.
In the following sections, we will successively question the relevance of each one of those hypotheses. Our discussion will rely on the tight connection that exists between BWE's approach and Pareto's definition of the *Homo economicus*: we will be able to highlight that BWE shares many methodological difficulties with Pareto — concerning the definition of one's true preferences — but also that, unlike Pareto, BWE confers a normative value to preference satisfaction that cannot be justified within Pareto's approach.

So as to highlight the possible difficulties of BWE, we will consider:

> **Petula's choice of train**: Petula, a pessimistic individual, must take the train to go to her workplace, and has a correspondence halfway (the second train leaves at 8h40). She knows that the first journey takes 30 minutes in normal conditions, but also that delays may happen (incidents actually occur on 1% of the journeys). She has the choice between taking the 7h and the 8h train. If she takes the 7h train, then she is almost certain not to miss her correspondence, even if an incident occurs. If there is no incident, she will however arrive too early at the second station, and is likely to regret not having slept an additional hour. If she takes the 8h train, then she will arrive on time for the second train if there is no incident, but she is likely to miss it in case of an incident. Her pessimism inclines Petula to systematically take the 7h train.
>
> Suppose also that Norbert, the neoclassical agent, faces the exact same problem than Petula. After having carefully thought about the pros and cons of taking the 8h train, Norbert decides to systematically take the 8h train. Although he sometimes misses his correspondence, this is compensated by the additional sleeping hour he gets all the other days.

From the social planner's perspective, it seems that Petula is making a mistake when taking the early train (she may believe that the probability of an incident is much higher for instance), since Norbert — Petula's neoclassical *alter ego* — takes the later train. The social planner should therefore nudge Petula to take the 8h train: this is possible by informing more explicitly that incidents are relatively rare events, or also by offering in a single package the ticket of the 8h train with the ticket of the second train. This kind of policies could induce Petula to take the 8h train, and therefore to better satisfy her true preferences. We therefore assumed that (i) Petula's preferences can be purified from her pessimism, revealing her true preferences (and therefore that it makes sense to imagine the behaviour of Petula's neoclassical *alter ego*, Norbert), (ii) it is normatively desirable to satisfy Petula's true preferences, and (iii) it is possible to elicit Petula's true preferences (and therefore to identify what Norbert would do).

## 3.3 Logical actions and the *Homo psychologicus*

In this section, we highlight the similarities between BWE and Pareto's study of logical actions: we will then be able to highlight a fundamental difficulty within Pareto's approach that may seriously undermine hypothesis (i), according to which the individuals are defined by underlying coherent preferences.

BWE assumes that individuals want to improve their well-being, as judged by themselves: we find here the first criterion of a logical action, the intention of satisfying a subjective purpose. It is then suggested that individuals often make bad choices (either due to information issues or to a lack of self-control), and that the choice architect can help them to achieve their own goals. To reuse the definition of Pareto, we can say that individuals often make actions which logically conjoin means to ends from their own standpoint, but not from the standpoint of the choice architect (this idea is explicitly stated by Bernheim and Rangel (2009, p.55)), a person who is supposed to have a more extensive knowledge. For instance, in the case of Luke's choice, we assumed that Claire knew how the choice architecture could influence Luke's choice, since she already offered similar contracts to other individuals, and therefore noticed the behavioural pattern induced by a specific choice architecture. Libertarian paternalism is therefore grounded on the will of creating a choice architecture such that the individuals perform *in fine* logical actions.

We have seen that BWE relies on a dual nature of the individual, with on the

one hand an optimising self, and on the other hand a psychological self. Similarly to Pareto who considers that the human is an aggregate of different kind of *Homines*, BWE pictures the individual as the aggregation of a *Homo economicus* and a *Homo psychologicus*. The *Homo economicus* describes the behaviour of the individual when she takes into account all the relevant elements for her choice (she therefore performs a logical action), while the *Homo psychologicus* describes the behaviour of the individual when she is not optimising, by following for instance simple heuristics. BWE's central claim is that individual choice is more accurately described by the *Homo psychologicus*, while the individual prefers to choose as the *Homo economicus* would choose. Proponents of libertarian paternalism argue that choice architects should design the choice architecture such that individual choices are the result of the choice of the *Homo psychologicus* (which are typically not cognitively demanding), while ensuring that the resulting choice coincides with what the *Homo economicus* would have chosen.

We can however highlight a fundamental ambiguity within this reductionist approach: Pareto did not indeed precise whether the different *Homines* of the individual simply *complete* or *change* the behaviour of the *Homo economicus*. Demeulenaere (1996, pp.175-177) gives the following example: consider a Muslim who wants to buy a prayer mat; the religious dimension of the transaction does not modify the economic interest for the individual to purchase her good at the lowest price. The *Homo religiosus* simply completes the *Homo economicus* by defining her ends. Now consider the same Muslim: the respect of her religious principles does not allow her to lend her money in order to get an interest. There is in this second situation a clear opposition between the behaviour of the *Homo economicus* and the one of the *Homo religiosus*. There are therefore two incompatible understandings of Pareto's approach: we can either consider the phase of aggregation of the different dimensions of the individual as the determination of the true preferences of the individual — that will give an objective to the instrumental *Homo economicus* — or as the addition of constraints which may enter in contradiction with the objective of the *Homo economicus*.
A major difficulty of BWE is that the very same problem arises: while the proponents of this model usually see the *Homo psychologicus* as inducing choices that do not satisfy the true preferences of the *Homo economicus*, the question of whether the *Homo psychologicus* may also determine the objective of the *Homo economicus* is not tackled. Consider Petula's choice of train. Being pessimistic may induce her to hold false beliefs about the likelihood of an incident. Her pessimism may also make her strongly averse to risky situations in general (since

she is convinced that bad events are very likely to occur). Unlike Petula, Norbert is not pessimistic and objectively treats the probability that bad events occur: but is Norbert still risk averse? A difficulty arises because Petula's risk aversion is linked to her psychological perception of risky choices: removing pessimism from Petula's preferences may also change her true preferences, i.e. what she would have done if she did not have a false belief about the probability of an incident. By defining Norbert as Petula's neoclassical *alter ego*, two distinct operations were possible: we could either consider that Norbert is just like Petula, but without false beliefs, or also that Petula's extreme risk aversion is due to her pessimism, and is therefore not relevant for Norbert's evaluation of the different alternatives. In this second situation, Norbert is not risk averse any more, although risk aversion seems to be a reasonable component of neoclassical preferences. Being pessimistic can indeed mean that people tend to perceive bad outcomes as more salient: but nothing is said whether there is a 'rational' level of risk aversion or not. We could for instance argue that rational players should be risk neutral: being risk averse can for instance be the result of the tendency to frame choice problems in terms of gains, and therefore be the result of loss aversion[1]. We will treat this point in more details in the case of intertemporal choices in section 3.6, by discussing the discount factor Oscar ought to use if he were rational.

In this situation, considering the behaviour of the *Homo economicus* independently of the behaviour of the *Homo psychologicus* is problematic: the preferences on which the *Homo economicus* would act are indeed partially defined by the psychological biases that prevent this same *Homo economicus* from satisfying them. Assuming the existence of true preferences therefore requires that the different *Homines* constitutive of the individual — in a reductionist perspective — are substitute rather than complementary.

The first methodological issue of BWE is the implicit assumption that the individual is nothing else than an instrumental entity, whose nature is comparable to a computer. The individual is indeed programmed to act in a predetermined way (satisfying one's true preferences), and is sometimes 'defective', in the sense that the action is not undertaken by the instrumentally rational *Homo economicus*, but by the psychological *Homo psychologicus*. As underlined by Georgescu-Roegen (1971), this computational perspective of the economic agent was already present

[1] We can indeed justify the choice of riskless prospects as the expression of loss aversion when the prospects are framed in terms of gains: when Luke must choose between €20 and €60 with probability 1/3, his reference point can be €20 rather than €0. Luke therefore chooses the riskless prospect because he wants to avoid the possibility of losing €20 with probability 2/3.

in Pareto's analysis:

> As Pareto overtly claimed, once we have determined the means at the disposal of the individual and obtained a 'photograph' of his tastes ... the individual may disappear.' The individual is thus reduced to a mere subscript of the ophelimity function $\phi_i(X)$. The logic is perfect: man is not an economic agent simply because there is no economic process. There is only a jigsaw puzzle of fitting given means to given ends, which requires a computer not an agent (Georgescu-Roegen, 1971, p.343)

BWE is more descriptively accurate than neoclassical economics since the *Homo psychologicus* offers a more realistic description of actual behaviour than the *Homo economicus*. However, the individual is still reduced to a computer, programmed to satisfy her true preferences (her 'tastes' for Pareto). The difficulty of BWE is that nothing is said about the origins of such true preferences: the *Homo economicus* is indeed only defined as an entity satisfying exogenous preferences, and the *Homo psychologicus* is only considered as another instrumental entity, whose procedure of choice is inferior to *Homo economicus*'s one. Within this model, one's true preferences must therefore be a-psychological, i.e. the psychological processes that affect our decision-making process must not affect our true preferences.
Accepting hypothesis (i), according to which an individual is defined by underlying coherent preferences, therefore requires offering a purely reason-based account of one's true preferences. We however already highlighted the difficulties of this approach in chapter 1, and in particular that it is generally not possible to define unambiguously what would be the true preferences of the individual.

## 3.4 Should I be rational?

Suppose however that hypothesis (i) is true, and therefore that human psychology can be reduced to reasoning imperfections without impact on the ends the *Homo economicus* intends to pursue. The question that follows is to know whether the satisfaction of those true preferences constitutes a valid normative criterion. Although this assumption seems relatively natural, several arguments may seriously undermine this claim: (a) non-logical actions are not necessarily mistakes, (b) *Homo economicus*'s behaviour can be self-defeating, and (c) in a welfarist perspective, we can argue that the feelings expressed by the *Homo psychologicus*, although not rationally grounded, should be taken into account.

It is worth noticing that Pareto only considered the *Homo economicus* as a descriptive model of human behaviour (empirically valid in a few settings such as repeated markets (Plott, 1996) and a useful abstraction for the study of markets (Pareto, 1909, chap.3, §§65-66 and §87). There is in particular no value judgement about the normativity of coherent preferences: in his attempt to study human actions, Pareto only suggested that there exists a specific set of actions (that occur for instance in repeated markets) for which the individuals tends to be instrumentally rational. Behavioural economics does not particularly contradict this statement, since it only shows that in many situations within the field of study of economics (which is larger than repeated markets) people are usually not rational, in the sense that their revealed preferences are incoherent — Pareto himself would probably have agreed with this result, a huge part of his work consisting in the study of non-logical actions. An interesting phenomenon is that the 'anomalies' highlighted by behavioural findings seem to have been understood as 'anomalies of behaviour' rather than 'anomalies of the theory': it is therefore not the theory according to which individuals are defined by coherent preferences that is wrong, but the individuals who make mistakes, i.e. whose behaviour is 'wrong' with respect to a norm of behaviour. This is for instance the case in Kahneman and Tversky's paper on prospect theory (Kahneman and Tversky, 1979): they are indeed talking about anomalies of *preferences* (p.275, p.277), suggesting implicitly that 'normal' preferences should respect the axioms of rational choice theory. Coherent preferences are therefore the norm, while incoherent preferences are seen as an anomaly that should be corrected so as to provide sound normative assessments.

The assumption that individuals *should* be rational, although central in BWE, is however never properly justified. We acknowledge that in many settings, the satisfaction of one's true preferences is uncontroversially the right normative criterion: this is for instance the case when individuals must choose an option among identical offers with complex pricing systems. The true interest of the individual is to choose the option with the lowest price, although her bounded rationality may induce her to choose a costly option (there are in those situations an objective scale allowing the external observer to measure the true preferences of the individual[2]). There however also exist many cases in which it does not seem reasonable to expect from the individuals to act rationally. Consider for instance:

[2]We will argue in chapter 4 that the notion of an 'external observer' is problematic in normative economics, and then offer an alternative definition of an 'objective' scale without reference to this kind of observer.

> **Sports gambling**: Sarah is a supporter of Palaiseau soccer team. While she enjoys seeing her team win, she is also likely to feel bad when her team loses. Gambling can here be seen as a form of insurance: by betting against her own team, Sarah will obtain a financial compensation when the team loses, i.e. the emotional damage she will endure will be partially compensated by a strictly positive amount of money. In order to satisfy her true preferences (which are affected by the results of her team but also by her monetary gains), we should therefore nudge Sarah to bet against her own team.

Sports gambling is typically a situation in which an individual may become addicted and lose large amounts of money — justifying then paternalistic interventions so as to protect the individual against her addiction. In the above example, it seems however that gambling can actually be in our own interest (as a form of insurance). It seems however quite implausible that sports fans — due to the very nature of their preferences — would accept a policy inducing them to bet against their own team (they are indeed more likely to bet for their own team, although it is probably not rational).

Furthermore, non-logical actions, i.e. actions that are not undertaken by the *Homo economicus*, are not necessarily mistakes. In particular, it is possible that some of the actions we undertake in a social context are not individually but collectively rational: although we do not satisfy our true preferences, such social behaviours are relevant to the extent that they allow us to satisfy *in fine* our true preferences. Consider for instance the common claim that individuals do not save enough for their retirement, and therefore that they should be nudged so as to increase their level of savings (Thaler and Benartzi, 2004). Although the true interest of the individual would be to unilaterally increase her level of savings, it is not certain that simultaneously increasing the savings of all the individuals is desirable: this may indeed lead to a relatively low level of consumption at the aggregate level, leading to a macroeconomic crisis of under-consumption. A reason why Americans do not save a lot for their retirement may be that it is collectively rational to do so, since this ensures a high level of consumption and therefore a high level of economic activity. The satisfaction of one's true preferences also crucially depends on the behaviour of the other individuals: although it is in my interest to choose an option that better satisfies my true preferences *ceteris paribus*, this does not mean that this is still true if we have to nudge simultaneously all the individuals.

Another argument against the normativity of the satisfaction of one's true pref-

erences is that *Homo economicus*'s behaviour can be indirectly self-defeating (Parfit, 1984). This is typically the case in coordination games (such as the Hi-Lo game, Sugden (1991)) and games of commitment (such as the Toxin Puzzle, Kavka (1983)). The counter-intuitive implication of this observation is that, in those specific games, if an individual wants to satisfy her true preferences, then it is in her interest to adopt an apparently non rational behaviour: being irrational can therefore be rational in those games. Consider for instance:

> **Oscar's firm**: Oscar is running a firm that produces a good $X$, and is in competition with other similar firms (the situation is therefore a Cournot oligopoly). Oscar is convinced (without good reasons) that the demand in his sector will be very high next year. He therefore intends to produce a lot (more than what he would have produced at Nash equilibrium). Suppose that Oscar's rivals are not misled by optimistic expectations, and know that Oscar irrationally expects a high demand. Since they know that Oscar intends to produce more than at Nash equilibrium, their best reply is then to decrease their production (so as to avoid a situation of overproduction and a too low price).

Oscar is therefore producing more than his rivals: it is then quite possible that he gets a higher profit than at the Nash equilibrium (the intuition behind this result is that the resulting equilibrium can be close to the Stackelberg equilibrium in which Oscar is the leader and his rivals are the followers). By satisfying preferences that differed from his true preferences, Oscar managed to get a strictly higher profit: being optimistic therefore gave him a strategic advantage. More generally, we will show in the second part of this thesis that, except for a very restricted range of games such as zero-sum games, payoff maximising behaviours (i.e. *Homo economicus*'s behaviour) is indirectly self-defeating.

The last argument against the claim that satisfying one's true preferences is a valid normative criterion has been exposed by Loewenstein and O'Donoghue (2004), when discussing the normative implications of their dual-self model. Although they suggest that satisfying the preferences of the deliberative self (i.e. *Homo economicus*'s preferences) is a defensible criterion — since 'it represents how people would 'like' to behave' — they also argue that individual welfare also depends on our affective self (i.e. *Homo psychologicus*'s preferences). They illustrate their argument with the choice between driving and flying: although I know that flying is safer than driving, I am also likely to experience a higher fear when flying.

Although this fear is not rationally grounded, I still actually feel it, and this should be taken into account in welfare analysis. This argument is applicable to Petula's choice (the two others arguments are indeed less relevant for Petula): suppose that Petula has been successfully nudged, and takes the 8h train. Although it was in her best interest to take this train, she is likely to be worried during the whole journey: this makes her travel uncomfortable, and should therefore be considered as an additional cost for her — cost that Norbert does not support.

The second methodological issue of BWE is that hypothesis (ii), according to which the satisfaction of the true preferences constitutes the normative criterion, is not properly justified. It is probably only the result of the progressive recognition of Pareto's *Homo economicus* as a normative model of behaviour: economics was indeed defined as the science of logical actions, for which the individuals are rational and manage to satisfy their true preferences. If an individual is not rational in a situation studied by economics, for which she is supposed to be rational, then her behaviour is an anomaly. The progressive extension of the scope of economic analysis to situations in which the players do not perform logical actions led economists to the conclusion that the wedge between theoretical predictions and empirical observations was the result of anomalies of behaviour rather than anomalies of the theory. We will discuss in more depth in chapter 4 some philosophical difficulties of the preference satisfaction criterion.

## 3.5 Eliciting one's true preferences

Suppose now that hypotheses (i) and (ii) are true. We now investigate the third hypothesis of BWE, i.e. whether it is possible to elicit the true preferences of the individual when such true preferences exist. We know that individuals' revealed preferences may not correspond to their true preferences, since they may depend on the choice architecture. The first possible approach to elicit the true preferences of the individual would be to directly deduce them from the set of revealed preferences. This approach is for instance endorsed by Bleichrodt et al. (2001), who assume that the individuals present loss aversion, and try then to deduce the true and unbiased preferences of the individuals from their stated preferences. Rubinstein and Salant (2012) adopt a similar approach and try to deduce from 'behavioural data sets' the true preferences of the individual. However, we cannot know *a priori* the list of the different biases that influence individuals' decisions. In the case of Bleichrodt et al. (2001), it can be doubtful to assume that the individuals only present loss aversion,

since it is not necessarily their only psychological bias, and it is not certain that they really suffer from loss aversion.
Knowing the actual preferences is therefore probably insufficient to obtain the true preferences, since there is an issue of identification within the determinants of behaviour between the true preferences and the possible decision flaws that affect the choices of the individual. Although we can identify coherent preferences after removing some biases, we cannot be sure that those preferences are not the conjunction of the true underlying preferences and another bias. Although BWE explains individual choice as a combination of preference satisfaction and mistake (see for instance Köszegi and Rabin (2008), according to whom one's behaviour results from one's 'combination of preferences and errors' (p.1822)), the theoretical status of those 'mistakes' is not well identified. They are indeed understood as the deviation from the satisfaction of one's true preferences, but those true preferences can only be discovered one's we have purified our revealed preferences. We therefore cannot know whether people's revealed preferences are mainly composed of true preferences (e.g. a present bias), or of errors (e.g. discounting one's future utilities — see section 3.6 on this point).

Since it seems difficult to directly elicit the true preferences from an external standpoint, an alternative solution would be to help the individuals to discover their own true preferences: it would then be possible to extrapolate the preferences observed in this specific setting to other situations, since hypothesis (i) implies that the true preferences are context independent (if we observe the true preferences of an individual in a specific context, then those true preferences should remain the same in other contexts). If (i) is true, then we only need to design a framework in which the individuals tend to perform logical actions: we will then be able to discover their true preferences. This assumption is for instance implicit in many experimental works, in which individual risk aversion is controlled by a questionnaire (it is assumed that the risk aversion revealed in the questionnaire remains stable when the individual participates in the experiment).
Several authors — including Pareto (1909, chap.3, §1) — suggest that the discovery of the true preferences is the product of learning thanks to the repetition of the situation of choice. A choice architect could therefore perform repeated experiments in order to elicit the true preferences of the individuals. However, since the true preferences of the individual are defined subjectively, it is not possible to know at which moment the individual is satisfying them: there is indeed still an issue of identification, since it may not be possible to distinguish between the true preferences and a systematic decision flaw. There is therefore here the temptation of defining

*objectively* the ends of the individual (for instance as selfish ones), and to consider that the individual has achieved this state of rationality and performs logical actions if and only if she is satisfying the preferences expected by the experimenter. This is for instance what Binmore is doing in his analysis of the ultimatum game:

> Novices offer a fair amount because this is what their currently operative social norm recommends. Novices who are offered unfairly small amounts are programmed to feel resentful and so want to punish the proposer by refusing. But this behaviour changes over time as people dimly perceive that the norm they are using is not adapted to the problem with which they are faced. In the Ultimatum Game, people learn that it does not make much sense to get angry if offered too little, but the mavericks who initially make small offers learn much faster that it does not make sense to demand too much if one is nearly always refused. (Binmore, 1999, p.F22)

Binmore considers that the individuals tend to a payoff maximising behaviour in a repeated ultimatum game (since 'people learn that it does not make much sense to get angry if offered too little'), even though we cannot directly observe it since 'the mavericks who initially make small offers learn *much faster* that it does not make sense to demand too much if one is nearly always refused' (the individuals therefore do not converge to the Nash equilibrium). There is therefore here the implicit assumption that the individuals respect a social norm because they want to maximise their payoff, although we cannot know what the true motives of the individuals are: we can for instance consider that an individual respects a specific norm by conformism (see for instance the famous experiment of Asch (1955)), or — as suggested by Binmore himself (p. F19) — that the subjects want to achieve what they perceive as the experimenter's objective, since this one can be seen as an authoritative figure (Milgram, 1975).

Since we cannot make a clear distinction between the ends of the individual and the different factors that can influence her decision, it seems quite difficult to design an experiment for which 'the time allowed for trial-and-error adjustment is "sufficient" '. An apparent stable behaviour can indeed correspond to the pursuit of a specific end plus a systematic decision flaw. In the previous example, we can for instance assume that the true objective of an individual is to offer and accept only equal shares, but that she prefers to follow an unfair rule that was implemented during the experiment by conformism. Although the actual behaviour of the individual is well predicted by the theory according to which the individuals want to maximise their payoff, the underlying reasons of her choice are more complex.

Furthermore, she is unable to satisfy her true preferences.
Recall the situation of Claire and Luke: thanks to past interviews, Claire believes that Luke truly prefers risky assets. Under the assumption that Luke's true preferences remained stable over time (and therefore that their past interviews can give to Claire some meaningful indications on Luke's *current* true preferences), Claire cannot however be certain that Luke is truly a risk-lover. Suppose for instance that the satisfaction of Luke's true preferences would imply a risk neutral behaviour, but also that, for an unspecified psychological reason, he *systematically* slightly underestimates his potential losses. In this situation, although her revealed preferences tend to highlight underlying risk-seeking preferences (although he is still subject to framing effects, and therefore chooses riskless options when the choice problem is framed in terms of gains), Claire is unable to distinguish between this systematic mistake and Luke's underlying true preferences.

The third difficulty of BWE is that, although we assumed that the individual is defined by a unique underlying coherent preference ordering, it does not seem possible to distinguish between those true preferences and a systematic decision flaw. The crucial issue is that the social planner should be able to clearly define what would be the choice of the individuals if they were able to perform logical actions. Logical actions are however defined by conditions such that 'sufficient' repetitions and 'adequate' incentives: the qualification of 'logical' for a specific action is therefore subject to the personal interpretation of the observer. It seems therefore implausible to implement an impartial procedure that could isolate the true preferences from the actual preferences of the individual. The subjectivity of the social planner in her welfare assessment is also hinted within Bernheim and Rangel's framework: they indeed define a 'mistake' (the difference between actual behaviour and the satisfaction of one's true preferences) as 'a choice made in a *suspect* GCS that is contradicted by choices in *nonsuspect* GCSs' (Bernheim and Rangel, 2009, p.85, our emphasis), a GCS (generalized choice situation) being defined as the combination of a set of actions and a set of ancillary conditions. This issue of implementability is stressed for instance by Qizilbash (2012), according to whom LP implicitly relies on a too strong version of the informed desire view of welfare (it is not clear that any human being would be able to identify the cases in which one's true preferences are actually better satisfied). The risk is then that behavioural welfare economists impose their own normative views about the true preferences of the individual. A particularly salient case can be found in the very first sentences of Conly (2013):

> We are too fat, we are too much in debt, and we save too little for

> the future. This is no news — it is something that Americans hear almost every day. The question is what can be done about it. The most common answer is that, first, we should exhort ourselves to be better: we should remind one another that eating too much of the wrong thing will make our lives shorter and more painful; should write admonitory op eds about how our failures to save will cost us individually and as a society; should, generally, tell ourselves things that by and large we already know. Second, we should simply exert more willpower to make ourselves do what we have been persuaded is right. The trouble with these two strategies, and generally with attempts to bring about change through education and persuasion, is that they aren't very effective. In this book I recommend that we turn to a better approach, which is simply to save people from themselves by making certain courses of action illegal. We should, for example, ban cigarettes; ban trans-fats; require restaurants to reduce portion sizes to less elephantine dimensions; increase required savings, and control how much debt individuals can run up (Conly, 2013, p.1)

Our point here is not to discuss Conly's defence of coercive paternalism (in contrast with soft paternalistic approaches, such as libertarian paternalism), but her diagnosis of what is right and wrong about individual choices. Conly's central argument is indeed the same than BWE — due to their poor reasoning abilities, individuals are likely to make poor decisions in terms of their own well-being — but she defends a more radical approach to tackle those issues. It then seems that, although BWE's central claim is that people should be helped so as to satisfy their (subjective) true preferences, the satisfaction of those preferences should ensure that we do not have a 'too high' body mass index, we do not smoke, we save a relatively large amount of our incomes, and more generally that we give a high weight to the long term consequences of our actions. Although those positions seem reasonable, and that it can be argued that the reason why many individuals are actually obese or overburden with debt is their lack of self-control, we cannot be certain that an individual cannot truly privilege the short-term benefits of her actions to the long-term costs. If we accept that what fundamentally matters for our own well-being is the long-term consequences of our actions, then we could also argue that individual's true preferences should induce a vegetarian diet (since livestock farming is very costly in terms of natural resources, and therefore is against our interest in the long term), but also that individuals should stop using their cars... and it would make sense to argue that the reason why actual people are still eating meat or use

their cars is that they have a bias towards the present, and neglect the future costs of their actions. Although those preferences are perfectly defensible and reasonable, it seems dubious to argue that this is necessarily what we truly prefer. The risk is that economists impose *in fine* as our true preferences what they think it is rational to wish. Since the object of BWE is to incite the individuals to follow the choices that would have been done by their inner rational agent — which does not even exist — BWE implies a standard form of *ends* paternalism: the ends of the *actual* individual are indeed replaced by the ends of a *counterfactual* individual, the inner rational agent.

## 3.6 Oscar-case

Recall our introductory illustration of Oscar's savings choices: Oscar only saved a small proportion of his income when he was young, and now regrets his past choice. From the perspective of BWE, Oscar's initial choice was a mistake and therefore justifies a paternalistic intervention on young-Oscar such that retired-Oscar may benefit from a more comfortable pension. The object of this section is to question the claim that regretting one's choice reveals a past irrational behaviour: the notion is indeed central in BWE, since Oscar's regrets are supposed to highlight that he did not act in her best interest, and therefore that he would have preferred to be nudged. Our analysis of the Oscar-case will highlight the different methodological issues of BWE we discussed in this chapter.

### 3.6.1 Regrets and mistakes

Libertarian paternalists argue that the choice architect should nudge young-Oscar today, because young-Oscar is making a mistake when saving only a small proportion of his income. We can legitimately assume that retired-Oscar is likely to regret the choice of young-Oscar because, today, many individuals are in the situation of retired-Oscar, i.e. they did not save a lot for their retirement and now regret their past savings choices. 'Oscar' should therefore be seen as a statistically representative individual[3]. A nudge in this situation can be seen as a form of means paternalism if and only if young-Oscar would agree with the policy if he knew that he is likely to regret his choice later. We show in this section that this condition,

[3]We can notice that a few individuals may suffer from the nudge (the individuals who truly prefer to consume a lot today): the legitimacy of nudges may therefore be questioned in this situation, since nothing justifies *a priori* that such exceptional individuals can be sacrificed to the benefice of the greatest number. See Bovens (2009, p.211) on this point.

although quite intuitive, is true if and only if we accept a rather implausible model of individual identity.

The economic analysis of intertemporal choices assumes that the individual 'Oscar' is a set of transient selves representing a decision maker at different dates. For sake of clarity, we will denote by $t$-Oscar the self of Oscar at date $t$. Each $t$-Oscar has preferences over time-dependent outcomes $(x, n)$ — the promise to receive the outcome $x$ at date $(t+n)$. LP is grounded on the idea that $t$-Oscar can be characterised by two different types of preferences: his *true* preferences — the counterfactual preferences on which he would act if he had 'complete information, unlimited cognitive abilities, and no lack of self-control', whose associated utility function determines $t$-Oscar's welfare — and his revealed preferences — the preferences that are revealed by his choices (or equivalently, the preferences that determine his choices). LP's paternalistic claim relies on the assumption that there exists a discrepancy between young-Oscar's true and revealed preferences: young-Oscar should therefore be nudged such that he satisfies *in fine* his true preferences. Retired-Oscar's regrets are then taken as an indicator of the mistake of young-Oscar.
However, if retired-Oscar regrets the choice of young-Oscar, then it means that if *young*-Oscar had chosen differently, *retired*-Oscar would be better off. But under which conditions do we know that improving the well-being of retired-Oscar is actually in the interest of young-Oscar (and therefore that nudging young-Oscar on the basis that it will benefit retired-Oscar is actually in young-Oscar's interest)? We suggest that accepting the claim that retired-Oscar's regrets are a sufficient reason to nudge young-Oscar requires accepting the three following hypotheses:

(1) all the $t$-Oscars have the same true preferences,

(2) retired-Oscar's revealed preferences correspond to young-Oscar's true preferences,

(3) each self $t$-Oscar is rationally required to make time-consistent choices.

Those conditions (and the model of individual identity that supports them) are implicit in LP, and are not explicitly endorsed by its proponents: we nevertheless suggest that those conditions are necessary for LP's paternalistic claim, although they are both descriptively and normatively questionable. Those three conditions are actually tightly connected to the three hypotheses underlying BWE: (1) is indeed equivalent to (i), since the trans-temporal self is defined by a unique coherent preference ordering, (2) corresponds to (iii), since it is assumed that

we can unambiguously identify the true preferences of the individual, and (3) is equivalent to the normative claim of (ii), since Oscar ought to choose the option that satisfies his true preferences.

If condition (1) is not verified, then time inconsistency does not matter from the perspective of young-Oscar: the reason why retired-Oscar regrets young-Oscar's choice is simply that his true preferences changed over time — and not that young-Oscar's choice was irrational. Indeed, allowing for preference changes with ageing may justify that retired-Oscar's regrets are consistent with a rational choice of young-Oscar. The paternalistic claim therefore requires the stability of t-Oscar's true preferences over time[4].

Suppose therefore that condition (1) is verified. The reason why retired-Oscar disagrees with young-Oscar's choice is that they do not have the same *revealed* preferences, although they share the same *true* preferences. However, so as to be sure that young-Oscar would benefit from the nudge suggested by retired-Oscar, we need to assume that retired-Oscar's revealed preferences correspond to young-Oscar's true preferences (condition (2)): we should therefore assume that young-Oscar is mistaken, while retired-Oscar has a correct *ex post* assessment of young-Oscar's choice. We can for instance consider that young-Oscar's perception of his own interest is biased by a *present bias* (O'Donoghue and Rabin, 1999), and has a tendency to put relatively higher weights on immediate outcomes — this would be the reason why he chose to postpone his savings effort. *Hyperbolic discounting* — the tendency to care relatively less about the outcomes distant in time, as involved by a present bias — is indeed a well-documented phenomenon (Frederick et al., 2002), and would explain young-Oscar's time inconsistent choice.
The difficulty of this argument is that nothing justifies *a priori* the rationality of retired-Oscar — in particular if we assume that young-Oscar presents a present bias. It is therefore not certain that retired-Oscar's revealed preferences actually correspond to his true preferences (and therefore, by condition (1), to young-Oscar's true preferences). Retired-Oscar may for instance present a *bias towards the future* (Parfit, 1984, p.165). Suppose for instance that the satisfaction of young-Oscar's true preferences actually implied saving little for his retirement. When retired-Oscar is reminded of the pleasant life he had when he was young, he should

[4]We will discuss in chapter 4 (section 4.3.2) the possibility to justify paternalistic interventions if condition (1) is not verified, since the choice of $t$-Oscar can be seen as a choice involving several individuals (with different true preferences). This will however not be a case of means paternalism, and will not support the paternalistic claim of LP.

accept that his low pension is the legitimate cost for his past consumption. But if retired-Oscar is biased towards the future, then, when comparing the expectation of a future consumption of €100 with the memory of a past consumption of €100, he tends to prefer the future consumption of €100. What retired-Oscar really wants in this situation is to go back 40 years earlier to change his decisions, and then immediately enjoy the long term benefits of his choice 40 years later: the regrets he expresses today do not necessarily mean that his past consumption was a mistake, since the cost supported by young-Oscar is almost imperceptible from retired-Oscar's perspective. The effort that young-Oscar perceived as intolerable became retrospectively unimportant for retired-Oscar. In this situation, saying that retired-Oscar would have agreed with being nudged when he was young does not mean that it would have been true for young-Oscar. Nothing therefore justifies *a priori* the empirical validity of condition (2).

Suppose now that conditions (1) and (2) are descriptively accurate. The last condition implicitly stated by LP to ensure that retired-Oscar's regrets reveal the irrationality of young-Oscar's choice is that young-Oscar is rationally required to make time-consistent choices, i.e. to make choices with which all his future selves would agree. Unlike conditions (1) and (2), condition (3) is a normative rather than descriptive condition: it indeed states how $t$-Oscar is rationally required to discount his future utilities, under TS's conditions of unlimited cognitive abilities, perfect information and complete self-control. LP is indeed a 'prescriptive approach', i.e. it is an '[attempt] to offer advice on how people can improve their decision making and get closer to the *normative ideal*' (Thaler and Benartzi, 2004, p.S167, our emphasis). The exponential discounting model (Samuelson, 1937), which forbids time-inconsistent behaviours, is generally considered as a relevant normative model of intertemporal choices (and is implicitly considered as such by LP, since ir forbids time-inconsistent choices). This claim is for instance defended by O'Donoghue and Rabin (1999), according to whom time inconsistency leads to important welfare losses, and therefore that an exponential discounting may be preferable in terms of welfare (welfare is defined within their framework in a 'long-run perspective' (p.113), with an equal weighing of the utilities of each period). Furthermore, we can notice that, if condition (1) is true, then, under TS's conditions of perfect rationality and complete information, (3) is necessarily true. If $t_0$-Oscar is rational (in the sense that he ought to choose his action so as to satisfy his true preferences), then he should take the exact same decision when comparing two outcomes in $t_0$ and $t_1$, or in $t_n$ and $t_{n+1}$. He indeed knows that, in $n$ periods, he will face the situation of choice he faces today in the first case. If we assume that Oscar prefers

€100 immediately to €100 in the future, then we can define a discount factor $\delta_{0,0} < 1$ such that $t_0$-Oscar is indifferent between €100*$\delta_{0,0}$ in $t_0$ and €100 in $t_1$ (a discount factor $\delta_{t,n}$ should be read as the discount factor used by $t$-Oscar when comparing two outcomes at dates $(t+n)$ and $(t+n+1)$). Hyperbolic discounting means that $t_0$-Oscar tends to give a relatively higher value to immediate rewards or costs than more distant ones: the discount factor $\delta_{0,n}$ between two dates $t_n$ and $t_{n+1}$ therefore increases when $n$ increases, i.e. when the delay between the choice and its realisation increases. $t_0$-Oscar therefore chooses his level of savings as if he believed that he would be more patient in the future than he is today: but since Oscar's true preferences remain stable over time (condition (1)), then retired-Oscar will regret the choice of young-Oscar, since he is exactly as patient as were young-Oscar. Hyperbolic discounting therefore implies that young-Oscar will take decisions that retired-Oscar will regret: this therefore justifies a paternalistic intervention on young-Oscar in his own interest.

We have shown that the argument according to which retired-Oscar's regrets are a sufficient reason to nudge young-Oscar means that (1) those regrets are meaningful for young-Oscar, and therefore retired-Oscar and young-Oscar have the same true preferences, (2) retired-Oscar retrospectively knows what was in young-Oscar's best interest, and (3) young-Oscar should have used a constant discount factor when choosing his level of savings. Accepting LP's argument therefore implies that we should also accept the idea that Oscar is simply a set of transient selves $t$-Oscar with constant true preferences over time. This picture of Oscar's identity seems quite implausible, since it imposes the stability of individual true preferences over time-dependent outcomes (from a descriptive perspective), as well as the superiority of the exponential discounting model (from a normative perspective). We now show that an alternative account of Oscar's identity does not necessarily support the normativity of the exponential discounting model, and that it is possible to rationalise time inconsistent behaviours (nudging young-Oscar would therefore not be a case of means paternalism any more).

### 3.6.2 Does time inconsistency matter?

The object of this section is to investigate whether $t$-Oscar is rationally required to discount his future utilities with a constant discount factor or not. We should firstly notice that, although people actually discount future utilities, it is not clear whether they are rationally required to do so (and if not, whether discounting one's future utilities is rational or not). Indeed, if we assume that Oscar's objective is

that his life goes as well as possible, as a whole, then it is not certain that there is any decisive argument for discounting future utilities (see for instance Sidgwick (1874), Rawls (1971), Elster (1986), Broome (1991b)). Frederick (2003) notes that temporal neutrality (the claim that a person should give the same weight to all utilities, regardless of their temporal position) implicitly assumes 'that all parts of one's future are equally parts of oneself; that there is a single, enduring, irreducible entity to whom all future utility can be ascribed' (p.90). If we accept the existence of such an irreducible entity, then $t$-Oscar should not discount his future utilities.
By opposition to this 'simple' view of identity, Parfit (1984), among others, offers a 'complex' view, according to which such an irreducible entity does not exist: a person is a sequence of overlapping selves who are connected by different physical and psychological properties. Parfit suggests that there exists a relation of 'psychological continuity' (p.206) between $t_0$-Oscar and $t_n$-Oscar: there exist strong psychological connections between $t_0$-Oscar and $t_1$-Oscar (such as shared memories, values, beliefs, desires...), as well as between $t_1$-Oscar and $t_2$-Oscar, until $t_{n-1}$-Oscar and $t_n$-Oscar. The relation of 'strong psychological connectedness' is however not necessarily transitive: although Oscar can consider himself as the same individual than the individual he was yesterday, and than the one he will be tomorrow, he may not consider himself as the same individual than the one he was 10 years ago, nor than the one he will be in 10 years. Although there is some physical and psychological continuity between $t_{-10}$-Oscar, $t_0$-Oscar, and $t_{10}$-Oscar, they are not necessarily the same person, since they do not necessarily share the same memories, values or preferences.

Parfit then argues that, from the standpoint of the decision maker, it is not personal identity but psychological connectedness that matters (Parfit, 1984, p.245): knowing that your personality is likely to evolve over time, you are not rationally required to care as much about your further future than your closer one (p.158). If we accept the complex view, then the different selves of Oscar can be seen as different persons: it is therefore not irrational for $t$-Oscar to discount his future utilities, since his future selves are not entirely himself. The idea that my future selves are not 'entirely' myself, although I may have a lot of common with them, can be captured by the notion of *psychological distance* between temporal selves. Psychological distance relates to the difficulty for people to experience the feelings and subjective states of others (either their future selves, Wilson and Gilbert (2003), or other individuals, Andersen and Ross (1984)). Liberman et al. (2007) define for instance 'psychologically distant things [...] are those that are not present in the direct experience of reality' (p.353): at date $t_0$, since $t_1$-Oscar does not exist yet,

$t_0$-Oscar cannot directly experience $t_1$-Oscar's utility. If $t_0$-Oscar is unable to fully experience the utilities of his future selves, then he is legitimated to discount their utilities. The existence of a non-null psychological distance between me and my future selves may therefore rationally justify the discounting of my future utilities. The question that follows is therefore the determination of a criterion to measure this psychological distance between temporal selves. Such a criterion may then give insights into how the individual is rationally required to discount future utilities. If we show that there exists a plausible measure of psychological distance such that hyperbolic discounting is not irrational, then the paternalistic claim of LP will be seriously undermined (the exponential discounting model would indeed not be the unique defensible normative model of intertemporal choice). Since the notion of psychological distance is related to the ability to experience the utilities of one's future selves, we suggest defining the psychological distance $0 \leq d_{0,n} \leq 1$ between $t_0$-Oscar and $t_n$-Oscar as the loss of welfare experienced by $t_0$-Oscar when he gives €1 to $t_n$-Oscar. The distance $d_{0,n}$ therefore captures the idea that $t_0$-Oscar is 'less' able to experience the utility of $t_n$-Oscar than his own utility.
We can now redefine the discount factor $\delta_{0,n}$ in terms of psychological distance. Recall that $t_0$-Oscar is indifferent between €100 in $t_{n+1}$ and €100*$\delta_{0,n}$ in $t_n$. The difference between those two outcomes can be interpreted as the cost supported by $t_0$-Oscar when $t_n$-Oscar gives €100 to $t_{n+1}$-Oscar. This cost can therefore simply be measured as the difference between what $t_0$-Oscar would experience if he gave €100 to $t_n$-Oscar — i.e. €100*$(1-d_{0,n})$ by definition of the psychological distance — and what he would experience if he gave €100 to $t_{n+1}$-Oscar — i.e. €100*$(1-d_{0,n+1})$. We have therefore:

$$(1-\delta_{0,n}) * 100 = (1-d_{0,n}) * 100 - (1-d_{0,n+1}) * 100, \tag{3.1}$$

$$\delta_{0,n} = 1 - (d_{0,n+1} - d_{0,n}). \tag{3.2}$$

From the perspective of $t_0$-Oscar, if two temporal selves are relatively close, then the discount factor $\delta_{0,n}$ tends to 1 ($t_0$-Oscar therefore assigns similar weights to selves who are relatively close, from his perspective).

Consider firstly LP's description of Oscar's identity. So as to ensure that condition (3) is verified, we must have a measure of the psychological distance between two successive selves such that the relative distance between selves $(d_{0,n+1} - d_{0,n})$ does not depend on $n$ (otherwise Oscar is likely to make time-inconsistent choices). Furthermore, we can notice that, if $t$-Oscar is rational and does not hold false

beliefs, then he knows that condition (1) is true. He could therefore perfectly anticipate the experience of his future selves: the psychological distance between the different selves $t$-Oscar is therefore necessarily null, implying that $\delta_{t,n} = 1$, for all $t$ and $n$. LP's conditions therefore imply that $t$-Oscar is rationally required to be temporally neutral[5].

Consider now Parfit's description of Oscar's identity as a sequence of strongly psychologically connected selves. Suppose that the psychological connectedness between two successive selves can be measured by a parameter $0 \leq \beta \leq 1$. We can for instance interpret $\beta$ as follows: while $t_0$-Oscar agrees with 100% of the choices he makes today (because they are motivated by preferences, values, desires that he considers as being his own), he cannot be sure to agree with more than $\beta$% of the choices made by $t_1$-Oscar, because a fraction $(1-\beta)$ of the choices made by $t_1$-Oscar are motivated by preferences, values or desires that $t_0$-Oscar does not recognise as being his own[6]. Since $t_0$-Oscar only benefits from €$\beta^n$ when €1 is given to $t_n$-Oscar, the psychological distance between $t_0$-Oscar and $t_n$-Oscar is $d_{0,n} = 1 - \beta^n$. We can then deduce the discount factor used by $t_0$-Oscar:

$$\delta_{0,n} = 1 - [(1 - \beta^{n+1}) - (1 - \beta^n)], \tag{3.3}$$

$$\delta_{0,n} = 1 - \beta^n(1 - \beta). \tag{3.4}$$

The discount factor $\delta_{0,n}$ increases with $n$: $t$-Oscar therefore discounts his future utilities as if he believed that his future selves would be more patient than him. The psychological distance (as perceived by $t_0$-Oscar) between selves indeed tends to diminish as $n$ increases. From the perspective of $t_0$-Oscar, $t_{40}$-Oscar and $t_{41}$-Oscar are for instance almost the same person — a person who is quite different from $t_0$-Oscar. It therefore does not cost anything for $t_0$-Oscar to impose an important

[5] A solution would be to consider $t_0$-Oscar's probability of dying before $t_n$: since there is a probability that $t_n$-Oscar does not exist, $t_0$-Oscar would then be able to discount the utility of his future selves. This argument cannot however ensure that the difference $(d_{0,n+1} - d_{0,n})$ remains constant over time: this would indeed require that Oscar has the same probability of dying at each period, which is highly implausible.

[6] Frederick (2003) stresses the difficulty of defining an objective measure of psychological connectedness. He for instance measures $\beta$ by asking to subjects in an experiment to 'rate how similar you expect to be in the future compared to how you are now, and how similar you were in the past compared to how you are now. By similar, I mean characteristics such as personality, temperament, likes and dislikes, beliefs, values, ambitions, goals, ideals, etc.' on a scale from 0 (completely different) to 100 (exactly the same).

effort on $t_{40}$-Oscar to the benefice of $t_{41}$-Oscar.

Reconsider now the decision faced by young-Oscar a bit differently. Oscar is hired for his first job and must decide how much to save for his retirement. He also cares about poverty in the third world and intends to give a part of his salary to a charity. He also knows that he is likely to lose his charitable aspirations while getting older and wealthier. He therefore consciously decides to save a smaller proportion of his current income to be able to give more to the charity today, and imposes on his future selves greater savings efforts, since he cares less about his further selves with whom he does not identify. Within this context, the choice of young-Oscar is not irrational any more: the discount factor he applies to weight his future selves is indeed not a discount factor with respect to time, but a discount factor with respect to psychological connectedness. This implies in particular that the claim that retired-Oscar will regret young-Oscar's choices is not a sufficient reason to nudge young-Oscar today, since it does not mean that young-Oscar's choice was irrational.
LP claims that we should prevent people from being irrational (this is precisely the objective of means paternalism): since it is possible to rationalise time inconsistent behaviours by considering that what matters for $t$-Oscar is not the satisfaction of some stable true preferences but his degree of psychological connectedness with his future selves, we cannot defend nudges on the basis that they help young-Oscar to make better choices for *himself*. Nudging young-Oscar is indeed likely to cause some harm to young-Oscar and to benefit retired-Oscar: the well-being of young-Oscar will however be increased if and only if he has sufficiently strong psychological connections with retired-Oscar.

LP and nudges are often justified by claiming that we should protect individuals from their own mistakes while respecting their subjectivity and freedom of choice. The justification of nudges in intertemporal choices (on the basis of personal regret) however requires accepting that (1) people are defined by stable true preferences, (2) they have a correct assessment of their past choices, and (3) they should use a constant discount factor when comparing future utilities. We suggested that conditions (1) and (2) are descriptively inaccurate, while the normative claim of condition (3) — when we assume that a person is unified by the existence of her stable true preferences — implies that the only rationalisable time preferences would be temporal neutrality. Indeed, although LP states that *dynamic inconsistency* is irrational (condition (3)), it does not state explicitly which preferences actually are rational:

it seems however that, under condition (1) and TS's conditions of perfect rationality and complete information, nothing justifies that a rational individual — who ought to choose his action so as to satisfy his true preferences — is allowed to discount his future utilities (the psychological distance between temporal selves is indeed null). This implies that, among the set of internally consistent preferences, only temporal neutrality is rational.

## 3.7 Conclusion

We have argued in this chapter that the three hypotheses underlying BWE are quite fragile. The existence of true preferences indeed requires accepting a quite implausible theory of human behaviour, grounded on principles of rational choice and free from any psychology. It also forces us to accept a reductionist account of human beings, and to treat psychological characteristics as alien elements to our true self. Furthermore, BWE imposes the satisfaction of those true preferences as the normative criterion, without proper justification: we indeed highlighted many situations in which satisfying one's true preferences does not constitute a self-evident normative criterion. Finally, it does not seem possible to elicit such true preferences from the revealed preferences of the individual, since disentangling between true preferences and mistakes necessitates at some point the subjective interpretation of the observer. We suggest that those different issues are directly related to Pareto's theory of the *Homo economicus*: Pareto had indeed the same difficulty than BWE concerning the definition of a notion of true preferences (at least when he considered defining the *Homo economicus* as the entity in charge of logical actions in general, without a specific restriction of motive), and his definition of logical actions does not provide an objective criterion allowing us to identify with certainty whether a specific action is logical or not.

The challenge of behavioural economics consists more probably in questioning the assumptions underlying BWE rather than merely questioning the 'efficiency' of the individual in satisfying her true preferences. Considering the neoclassical framework as a benchmark for modelling individual behaviour, and progressively integrating psychological factors into the model, may not constitute a viable approach for normative analysis. This approach indeed presupposes that an individual 'Oscar' is merely a preference relation, and her life simply consists in the more or less successful satisfaction of these preferences. These true preferences are necessarily exogenous and remain stable from his birth to his death: BWE models Oscar as if he was simply a computer (with some programming imperfections), programmed to

satisfy a certain number of predetermined objectives given by his true preferences. This model does not probably provide an acceptable account of personal identity, since it results from the model of the Paretian *Homo economicus*, which was defined as a representative agent for investigation purposes and the study of market equilibrium (Boianovsky, 2013). In particular, it was explicitly designed so as to avoid integrating within economic models the complexity of individual behaviours. It can here seem a bit ironic that, as behavioural economists, behavioural welfare economists accept such a model based on principles of rational choice and free from any psychology — defining a psychology-free entity was precisely the objective of Pareto, but he confessed that this model could not offer a precise comprehension of individual behaviour (Pareto, 1916, §36).

An implicit assumption supporting the optimisation-based approach for modelling boundedly rational individuals is that, by progressively introducing psychological factors into the model, we will reach *in fine* an accurate description of how people actually behave. This means that the only difference that exists between a computer and a Human is their power of calculus: the computer is taken as the model of efficiency the Human should tend to. Although this kind of approach is quite relevant for descriptive purposes (as defended by Rabin (2013)), such models cannot be used to derive normative prescriptions on individual welfare, since this would require the actual existence of neoclassical preferences, 'hidden' behind our psychology. Reconciling behavioural and normative economics requires leaving the assumptions of BWE: behavioural economics indeed does not tell us that Humans are defective computers, but rather than Humans are not computers.

Chapter 4

# Preference satisfaction and individual autonomy

## Contents




**Abstract:** We question in this chapter the welfarist claim of BWE, according to which the satisfaction of one's purified preferences matters. We argue that this claim implicitly assumes that individuals are simply passive *loci* of experience: BWE denies the status of *agents* for individuals, and in particular that individuals can actively contribute to the shaping of their own preferences. This assumption derives from the third-person perspective economists endorse when providing normative assessments: they indeed take the standpoint of an omniscient and omnipotent social planner, whose objective is to design the society such that the individuals achieve *in fine* their own ends. We argue on the contrary that behavioural findings, by questioning the idea of true preferences, invalidate this welfarist model. We then suggest a normative criterion in terms of individual autonomy, according to which it is the ability to choose and accept one's preferences that matters.

## 4.1 Introduction

A central argument of the liberal tradition in economics against paternalistic interventions in economic activities is that competitive markets are an institution in which individuals actuated by the pursuit of their own interest achieve unintentionally a socially desirable outcome. This argument — which can be traced back at least to Smith's idea of the invisible hand — is commonly associated, within the framework of neoclassical welfare economics, to the first fundamental theorem of welfare economics, according to which a competitive equilibrium is *Pareto-efficient*. Consider an economy composed of $N$ individuals with coherent preferences over the possible states of the world; a state of the world $S_1$ leads to a Pareto-efficient outcome if and only if there does not exist a state $S_2$ such that all the individuals prefer $S_2$ (strictly prefers for at least one individual). Since preference satisfaction matters, Pareto-efficiency is socially desirable because it is not possible to offer to an individual an outcome she prefers without causing some harm to another individual.
According to neoclassical welfare economics, markets are therefore successful in satisfying individual preferences: since preference satisfaction is the normative criterion, paternalistic interventions in markets should not be allowed. Behavioural economics however highlighted that individual preferences are generally incoherent, and therefore that — in a welfarist perspective — satisfying one's revealed preferences is not necessarily desirable. Behavioural welfare economics then legitimates paternalistic interventions in markets, since markets are not able to satisfy individual *true* preferences.

BWE therefore justifies paternalism on the ground that the satisfaction of one's true preferences matters, and that neither the individuals nor markets are able to satisfy the true preferences of the agents. We have shown in the previous chapters that the assumption that individuals are characterised by underlying coherent preferences is not properly justified: it is indeed probably a property of the representative agent of repeated markets rather than of a real agent, and it also requires the existence of a latent mode of reasoning able to generate counterfactual coherent preferences. The aim of this chapter is to question the second claim of BWE, according to which the satisfaction of those purified preferences — if such preferences exist — is the normative criterion. The guiding thought of this chapter is that normative economists see themselves as if they were in an 'objective' position, and therefore that their normative recommendations should be accepted by rational individuals: we argue here that the satisfaction of one's preferences

matters only if the agents have *chosen* those preferences, and therefore recognise those preferences as their own. Normative economists should therefore not try to impose their own normative values, and simply try to give to the individuals the means that will allow them to choose their own preferences. This position is relatively close to what Qizilbash (2011) labels the 'thick view' of Sen's capability approach (Sen, 1999, 2009), according to which public reasoning plays a central role in the definition of the collective preferences of the society: it is argued that theoretical reasoning cannot, on its own, provide a normative criterion for individuals with different views (Sen, 2004, p.78). It then falls to the citizens, as a matter of sovereignty, to decide collectively what matters for themselves.

We firstly describe the third-person perspective economists endorse when providing normative assessments, and show that this approach relies on a vision of individuals as passive entities, unable to build their own normative assessments (section 4.2). We then argue that acknowledging the status of autonomous agent of the individual seriously undermines BWE's paternalistic claim: we indeed suggest that the normative challenge raised by behavioural findings is that individual freedom may be restricted by decision biases, and subsequently defend a normative criterion of individual autonomy rather than preference satisfaction (section 4.3). We then argue that claiming that preference satisfaction matters may offer a distorted view of many societal issues, since it neglects the central role of democratic processes in the formation of individual preferences. We illustrate this point by reviewing the conditions under which common pool resources can be sustainably managed (section 4.4).

## 4.2 The market and the hive

In neoclassical economics, the theoretical framework of competitive markets provides an efficient mechanism to allocate resources and satisfy individual preferences at the aggregate level. It has therefore been elevated to the status of an *ideal*: economists then use this theoretical framework to detect deviations from the ideal outcome, and then define the policies that should be implemented to correct those deviations, so as to reach *in fine* the socially desirable competitive market equilibrium. Four kinds of situations are then often considered as *market failures*, i.e. situations in which the market equilibrium differs from the Pareto efficient competitive equilibrium: externalities (Pigou, 1920), market power, asymmetric information (Akerlof, 1970) and bounded rationality (Bennett et al., 2010, Sunstein,

2014b). Economic analysis then provides a large toolbox to the social planner to fix those market failures and restore the first best (or settle for the second best, Lipsey and Lancaster (1956)). We can mention for instance the Pigovian tax to internalise negative externalities, Lindhal prices (Lindhal, 1919) for the provision of public goods (which can be understood as a form of positive externality), preference revelation mechanisms (Vickrey, 1961) in the case of asymmetric information, and nudges for boundedly rational individuals (Thaler and Sunstein, 2008). The object of economics consists in designing different types of *incentives* (mainly monetary, but also psychological ones in the case of nudges) so as to overcome those market failures.

Grant (2002) suggests that the main reason why incentives seem to be ethically unproblematic is that they can be assimilated to a form of *trade*. Consider an individual (P1) who intends to make an option $A$ more attractive than the other options to another individual (P2). The incentive typically consists in a cash payment, such that the option $A$ becomes unambiguously more interesting from P2's perspective (P2 may therefore decide to choose $A$, although he would have chosen an option $B$ in the absence of incentive). P1 and P2 therefore reach an outcome that is beneficial for both of them, since P1 preferred that P2 chose $A$ rather than $B$, and P2 has a strictly higher payoff when choosing $A$ rather than $B$. An incentive therefore involves a voluntary action by all parties (P1 chose the level of the incentive, and P2 was not forced to choose the option $A$) leading to a mutually beneficial outcome: incentives are therefore ethically preferable to other forms of interventions such as coercion. Following the same line of argument, it can also be argued that nudges are ethically unproblematic. Suppose that P1 is a benevolent choice architect who knows that P2 truly prefers $A$ over $B$, but that P2 is likely to choose $B$ due to some psychological bias. Here again, the nudge allows the players to reach a mutually beneficial outcome (since P1 is benevolent, she indeed only cares about P2's welfare), without forcing P2 to choose $A$. Grant then highlights that, although economists commonly believe that their thinking about political economy is in the continuation of the Scottish and English Enlightenment (such as the idea of the invisible hand, which is associated to the first fundamental theorem of welfare economics), the word 'incentive' does not appear in any of their writings[1]. Grant suggests that this evolution is closely related to the emergence of behavioural psychology in the early 20th century, and in particular to the idea of scientific management developed by Taylor (1911), that deeply influenced the way economists perceive the functioning of a market society. This evolution can

[1]Grant simply mentions an anecdotal use by J.S. Mill in the *Principles of Political Economy*, and by Ricardo in the *Principles of Political Economy and Taxation*.

be captured by a fundamental shift in metaphor from Smith's 'invisible hand' and the British thinkers of the 17th and 18th century to the current figure of the 'social engineer': while the society was conceived as a huge *clock*, functioning automatically and predictably according to natural laws, it progressively became 'an amalgam of forces in constant flux that can be directed to bring about progress' (Grant, 2002, p.117). Since there is no reason *a priori* for the spontaneous order to be socially optimal, the social planner ought to implement the adequate incentives so as to steer individual behaviours into the right direction.

Let us suggest an alternative metaphor for representing how welfare economists (both neoclassical and behavioural) conceive human society, a giant *hive* designed by a *beekeeper*. A society is composed of a group of individuals interacting with each other within specific institutional rules. The only aim of an individual is to satisfy her preferences, and all her actions are guided towards the satisfaction of her preferences. Those preferences — as soon as they are coherent, as postulated in neoclassical welfare economics — can be represented by utility functions defined over the set of actions: similarly to bees whose only objective is to produce honey, the individuals' only objective is to produce utility. Similarly to the production of honey which is beneficial to the bees (as a source of food), the production of utility is beneficial to the individuals (as a source of welfare). Similarly to man-made beehives which are scientifically designed to increase the production of honey compared to the production of a natural beehive (allowing the share of the surplus between the bees for their own consumption and the beekeeper), it is possible to scientifically design the society such that the total production of utility is higher than at the initial natural order. Similarly to the beekeeper who can introduce frames with some honeycomb to ease the production of honey, the social planner can build markets to ease the processes of exchange and matching, and then the production of utility. A central element of this analogy is that individuals — in line with behaviourism — are *passive* entities, and are only responding to external *stimuli*, such as monetary incentives and nudges.

The economist is therefore looking at the society as a beekeeper is looking at a hive: her objective is to determine the optimal structure for the hive such that the total production is maximised. Since individuals are perfectly responsive to incentives and nudges, their actions can be guided by the planner so as to compensate possible market imperfections. By only altering the institutional environment of the hive (such as implementing taxes — that do not fundamentally change the

functioning of the market — or nudges), the individuals are not coerced, and can still refuse to choose the option the planner wants them to choose — if they want it. Economists — as social scientists — can then endorse this role of *social designer*, as clearly stated by Roth (1991):

> In the long term, the real test of our success will not be merely how well we understand the general principles which govern economic interactions, but how well we can bring this knowledge to bear on practical questions of microeconomic engineering, to design appropriate mechanisms for price formation [...], dispute resolution, executive compensation, market organisation, etc. [...] Just as chemical engineers *are called upon* not merely to understand the principles which govern chemical plants, but to design them, and just as physicians aim not merely to understand the biological causes of diseases, but its treatment and prevention, a measure of the success of microeconomics will be the extent to which it becomes the source of practical advice, solidly grounded in well tested theory, on designing the institutions through which we interact with one another. (Roth, 1991, p.113, our emphasis)

Economic theory — thanks to the fundamental theorems of welfare economics — claims to have scientifically shown that competitive market is the best way to coordinate individual actions. The 'duty' of economists, as social scientists, is now to apply their results to real societies, by designing markets and price mechanisms that will rule the interactions between individuals. Those interactions are not restricted to standard economic interactions (such as buying and selling goods), but include 'some of the most important markets that we are involved in — the matching markets that determine what schools we go to, what jobs we get, and maybe who we are married to' (Roth, 2012, p.343). Human societies can then be reduced to a vast market in which welfare is produced through exchange and matching: the ideal society economists want to design is therefore a complete market (for which the first fundamental theorem holds). The society should therefore be nothing more than a purely mechanical device designed to maximise the production of welfare, in which nothing else than efficiency and preference satisfaction should matter. This point is particularly salient in Roth's discussion of *repugnance*. Repugnance concerns certain types of transaction (organ markets) or activities (dwarf tossing) that are considered as morally unacceptable. Roth (2007) however considers that repugnance is similar in nature to a 'difficult technological barrier', i.e. to a constraint that should eventually be overcome so as to maximise social welfare:

> The persistence of repugnance in many markets doesn't mean that economists should give up on the important educational role of pointing to inefficiencies and tradeoffs and costs and benefits. But neither should economists expect such arguments to immediately win every debate. Being aware of the sources of repugnance can only help make such discussions more productive, not least because it can help separate the issues that are fundamentally empirical – like the degree of crowding out of altruistic donations that might result from different incentive schemes compared to how much new supply might be produced – from areas of disagreement that are not primarily empirical. (Roth, 2007)

Roth's claim is that empirical phenomena matter (crowding out effects for instance), but moral considerations and the disapproval of certain transactions on non-empirical basis (such as ethical or religious concerns) are not relevant. Economists should convince the population that considering a transaction as repugnant is pointless[2].

Economists see the world as if it was a giant hive, in which each individual seeks to maximise her individual welfare. They take the viewpoint of the beekeeper of this hive, who wants to maximise the total welfare produced within the hive, and who is also able to design the adequate incentives such that the individuals adopt the socially optimal behaviour. While producing policy recommendations, they are therefore endorsing the beekeeper position (so as to define the optimal policies), and then address their recommendations to the same abstract beekeeper. The interpretation of those policy recommendations are therefore 'if I were an omnipotent and omniscient planner, this is what I would do' (Sugden, 2013), or more precisely, 'if I were an omnipotent and omniscient planner, *and* if the real world was like the hive, *and* if what matters was preference satisfaction, then this is what I would do'.
The viewpoint from which normative assessments are made is the one of an impartially benevolent spectator (Sugden, 2013): the economist builds a model in which idealised agents interact according to predefined rules (individuals are actuated by an intrinsic motive of preference satisfaction, and 'guided' by different incentives), sets the model in motion, and then assesses the final outcome, from

[2]Note that in certain cases, the repugnance of certain activities seems to be incoherent with the acceptance of others, such as horse eating in California, which is forbidden for human consumption but not for pets (Roth, 2007, p.37). The cases of repugnance that Roth has in mind are however much more problematic from an ethical perspective, such as kidney exchange.

her position as a modeller. Therefore — unless her interests as an economist are directly related to the interests of a specific class of agents in her model (the ones funding her research for instance) — her assessment is impartial. Economists therefore adopt a *third-person perspective* so as to provide normative assessments.

Sugden (2008) argues that it is this peculiar, synoptic perspective of neoclassical and behavioural welfare economics that lead the proponents of BWE to argue that paternalism is inevitable, since 'economists are still inclined to think of normative analysis as a matter of solving optimisation problems using given data [...] about individuals' preferences. This leads them to conceive of incoherent preferences as a kind of corrupted data' (p.229). But if markets are understood as a means to create the opportunity for the individuals to make mutually advantageous exchanges, then it is still possible to recognise the normative properties of the market as a system of economic organisation, even if individual preferences are not coherent. This approach states that economic institutions should not be evaluated according to the degree to which preferences are satisfied, but rather in terms of the opportunities they provide to the individuals. This position is in the continuation of Buchanan (1968) normative analysis, according to which the normative problem that economists are facing is to determine fair agreements between individuals: rather than taking the standpoint of a 'benevolent despot' whose objective is the satisfaction of individual preferences, normative economists should directly advise the individuals of the forms of exchange that are likely to lead to mutually advantageous agreements. Sugden (2007) and McQuillin and Sugden (2012a) show for instance that competitive markets maximise the opportunities for the individuals to get what they want, even if their preferences are incoherent. Although we share the view that normative economics should not be addressed to an abstract social planner but directly to the individuals, and that preference incoherences do not necessarily justify paternalism, our criticism of BWE's paternalism will rest on a normative criterion of individual autonomy rather than opportunity.

## 4.3 Preference satisfaction and autonomy

We have seen in the previous section that preference incoherences may justify paternalism if and only if we accept that individuals can be reduced to passive utility producers. We firstly highlight that accepting this model implies that LP is not libertarian, since we cannot reasonably argue that individuals are still able to make free choices when they are nudged. We then argue that the normative issue raised

by behavioural findings is that people may lack of autonomy.

### 4.3.1 Why libertarian paternalism is not libertarian

Thaler and Sunstein (2008, pp.252-253) present LP as 'the real Third Way' between paternalism and libertarianism, since they argue that nudges can make people better off, as judged by themselves, without limiting their freedom of choice (LP is therefore a *means* paternalism as well as a *soft* paternalism). We already extensively discussed the limitations of LP's welfarist claim in the previous chapters, but we have not paid attention so far to LP's libertarian claim, according to which nudges preserve the freedom of choice of the individuals.

A central issue here is to precisely define what 'freedom of choice' means. Sunstein and Thaler evaluate the freedom of choice through the set of actions an individual has at her disposal – since they consider that nudges, when they are 'easy to avoid', do not restrict the freedom of choice of the individuals (2008, p.6) – and not through the set of actions within which an individual is *actually* able to choose her action. They are therefore focusing on a purely formal notion of freedom: our point is that it is not self-evident to argue that an individual subject to framing effects, and therefore whose choices are conditioned by frames, can make free choices. Indeed, Suntein and Thaler emphasise that nudges are not coercive, since the individuals can still choose another option than the one wanted by the choice architect *if they want it*. The difficulty of this argument is that real individuals, unlike the rational *Homo economicus*, are generally not able to choose what they truly want: this is precisely the reason why, according to Sunstein and Thaler, people should be nudged. As an illustration of this point, consider:

> **Rachel and Patrick's educational choice:**[3] Rachel and Patrick are two students with identical academic abilities, and must choose whether to apply to a university or not. Rachel comes from a relatively rich family and is used to live quite comfortably. If she does not go to university, she will not be able to keep her level of consumption and will probably difficultly accept her more modest life. On the contrary, Patrick grew up in a relatively poor family and is used to a modest standard of living. While he would greatly appreciate going to university so as to be able to get a high salary, he would not consider as a failure the option of quitting his studies before university.

[3]This discussion on the influence of social origins on educational choices is extracted from a paper written with Léonard Moulin, in which we study the potential segregative effects of the implementation of tuition fees in higher education, when the social aspiration of prospective students is biased by their social origins.

Due to their different social aspirations, it is likely that Rachel will be more tempted to pursue her studies than Patrick, since her desire to avoid downward social mobility is stronger than Patrick's desire for upward social mobility (Boudon, 1974, 1994, Breen and Goldthorpe, 1997). Although Rachel and Patrick have access to the exact set of opportunities, their different social origins imply a different perception of the available alternatives. Bourdieu (1974) for instance argues that students from a disadvantaged background will behave such that they will achieve what they perceive as an established fact: 'when you belong to a disadvantaged background, you cannot join University' (p.6). Patrick's modest social origins imply that he does not consider that the option 'going to University' is for him: although he is 'physically' able to go to University (if he decides to go to university for some reason, then this would technically be possible), he is 'psychologically' unable to do so.

We suggest therefore distinguishing between a set of *physically possible options* and a set of *psychologically possible options*: while the former corresponds to the set of options from which the inner rational agent could choose (as in standard decision theory), the latter isolates the set of options from which the individual, given her diverse decision biases and cognitive limitations, is actually able to choose an option. Patrick's social origins for instance cause him to self-censor, although the option is still available to Rachel. We want to argue here that one's freedom of choice should not be assessed based on a set of physically possible options (as in LP), but on the set of psychologically possible options. Suppose for instance that behavioural economists discover a specific frame such that, when the individual is not aware that she is subject to framing effects, she will systematically choose a specific option (a kind of default option for instance). Since the choice of the real individual is determined by the only choice architecture, can we still say that the existence of alternative options increase her freedom of choice? Indeed, although the set of physically possible options remains stable, the set of psychologically possible options shrinks to a single element: this situation is therefore equivalent, in terms of freedom of choice (in a substantive rather than formal sense), to a situation in which the only available action is the one chosen by the choice architect. The only difference is that the individual has the illusion of having a greater freedom of choice.

We define here freedom of choice in a negative sense (Berlin, 1958) as the *ability to choose without being coerced by elements external to one's self*, whether or not the choices of the self are determined or predictable. This definition of freedom would probably be compatible with Sunstein and Thaler's conception of LP, since it considers that individuals are free if they are not coerced by third parties, and if

their subjectivity (and therefore their true preferences) is respected. If we consider now the model of the inner rational agent retained by the proponents of LP, then the true self of the individual is her *Homo economicus*, and individual choices can be influenced by external factors such as psychological biases and framing effects (her *Homo psychologicus*). We can now show that this model of agency — necessary for the welfarist claim to hold — implies that the libertarian claim cannot hold.

Note that the welfarist claim holds only if nudges effectively improve individual welfare, by shaping the choices of the individual. If the choice architect is able to improve individual welfare thanks to framing effects, then it means that choices are not free: they are indeed conditioned by an element external to the self, the will of the choice architect. The mere influence of nudges on individual choices implies that the freedom of choice is not preserved, since the individuals are manipulated by the choice architect without their consent. This means that the libertarian claim cannot be verified if a planner wants to improve individual welfare by using framing effects. LP cannot be libertarian by construction, since it relies on the idea that the choice architect *should* manipulate the choice architecture so as to influence individual choices. If we accept the model of the inner rational agent — which is necessary to ensure that nudges can effectively improve our own well-being — then the libertarian claim can be verified if and only if people are *Homo economicus*, since their choices should not be influenced by the nudge without their consent. But if an individual is sufficiently rational not to be influenced by the nudge, then she is also sufficiently rational to satisfy her true preferences on her own.
Similar points have been raised by Hausman and Welch (2010) and Grüne-Yanoff (2012), who argue that, although Sunstein and Thaler free-choice condition seems to be verified, nudges interfere with individual autonomy. Hausman and Welch (2010, p.128) for instance argues that nudges 'exploit flaws in human decision-making to get individuals to choose one alternative rather than another', and that shaping individual preferences thanks to nudges violates the autonomy of the individual. Grüne-Yanoff (2012) stresses the incompatibility of the two claims of LP, by highlighting that the notion of welfare used by Sunstein and Thaler (as the satisfaction of counterfactual coherent preferences) is not compatible with liberal principles, such as the respect of the subjectivity and plurality of people's values. According to Grüne-Yanoff, nudges are manipulative since 'they deliberately circumvent people's rational reasoning and deliberating faculties, and instead seek to influence their choices through knowledge of the biases to which they are susceptible' (p.636). It is therefore because one's reasoning is deliberately influenced by the choice architect, i.e. that the will of the choice architect interferes with one's reasoning, that nudges

violate the autonomy of the individual.

We suggested previously that BWE was not able to identify the true preferences of the individuals — if such true preferences exist — and therefore that the choice architect is likely to impose to the individuals her own views about their well-being (such as preferring the fruit over the cake, or being temporally neutral). This suggests that, unlike Sunstein (2014b)'s claim that LP is a form of *means* and *soft* paternalism, LP will generally be a *ends* and a *hard* paternalism. LP indeed evaluates the freedom of choice of the individuals from the standpoint of individuals who are not subject to framing effects, and not from the standpoint of the real Humans who make choices (one's freedom of choice is evaluated through the set of physically rather than psychologically possible options). The choice architect indeed chooses *in fine* what matters for the individuals (since it is not possible to elicit with certainty a single coherent preference ordering), and imposes those preferences to the individuals by exploiting their limited freedom of choice.
Our last statement may seem a bit unfair, since there exist many situations in which nudges do not seem problematic at all. Imagine for instance that you must cross a street in London. Although you *know* that people drive on the left side on the road, you will not necessarily *spontaneously* look to your right before crossing (simply because you live in a country in which people drive on the right side, and you are used to this configuration of the road). It seems reasonable to consider here that being hit by a car simply because you did not look at the right side of the road is a mistake: there is indeed a relatively objective measure of your success in crossing the road, i.e. reaching safely the other side. In this situation, the signs 'Look left' and 'Look right' on the zebra crossing seem to be perfectly acceptable nudges, that are likely to prevent you from doing a serious mistake. Those kinds of nudges (including typically reminders and warnings) are often cited by the proponents of LP to defend the use of nudges, and in particular to highlight that many nudges are welfare-enhancing and not manipulative (see for instance Sunstein (2015) reply to Whitman and Rizzo (2015)). However, it is not because nudges are relevant in certain situations that we can justify their use in contexts in which defining the true interest of the individual is not straightforward. This is typically the case of retirement savings choices discussed in chapter 3: although BWE commonly claims that people who do not save a lot are mistaken (since they will eventually regret their choices), we argued previously that a more detailed analysis of intertemporal choices could seriously undermine this argument. The nudges we are criticising here are typically default options that induce people to choose an option that the choice architect has deliberatively chosen (even if she genuinely believes that it was in the

true interest of the nudgee). Our analysis will however still supports the use of a certain class of nudges, as long as they are designed so as to enhance the *autonomy* of the individual.

### 4.3.2 Preferences and autonomy

Consider the Oscar-case. Note that justifying imprudent behaviours (such as saving a small proportion of one's income) on the basis that they are not irrational does not imply that paternalism is not justifiable. Suppose for instance that a reason why young-Oscar prefers to consume more today is that he decides to start smoking. Although this behaviour can be rationalised because young-Oscar is not rationally required to care about retired-Oscar's health, we can argue that young-Oscar directly causes some harm to retired-Oscar (imprudent behaviours can therefore be *morally wrong* (Parfit, 1984, p.318)). If intertemporal choices can be seen as choices involving several individuals, we can appeal to Mill's harm principle to justify a paternalistic intervention on young-Oscar. A new difficulty then arises, viz. which self of Oscar should be privileged. The issue here is that modelling Oscar's choice as a game between multiple selves with their own preferences (that are nonetheless similar for two relatively close selves) does not provide a concept of welfare that could be applied to the individual as an enduring agent. In a multiple selves model, the determination of a normative criterion on welfarist grounds is therefore not straightforward, since we do not have at our disposal such a notion of enduring agent, and we have no reason *a priori* to privilege a self over another.
It should however be noticed that, unlike within Parfit's analysis of identity, changes in preferences, values, or desires can also be *initiated* by the agent, rather than merely *experienced*. Korsgaard (1989) for instance argues that persons are unified by the continuity of agency of their successive temporal selves, each of them being an active agent, contributing to the shaping of their own identity: what matters from this perspective is not psychological connectedness, but 'the view of myself as an agent, as one who chooses and lives a particular life' (p. 23). This argument relies on the Kantian position that we may view ourselves not only as objects of theoretical understanding (the passive *loci* of our experiences) but also as agents, as 'the thinkers of our thoughts, and the originators of our actions' (Korsgaard, 1989, p.18). What matters for Oscar is therefore the realisation of the life (in the sense of long-term commitments) he has chosen, as an autonomous agent, and not the experience of individuals with whom he is psychologically connected.

Suppose therefore that individuals are more than passive *loci* of welfarist

experience, in the sense that they are able to build their own identity, thanks to the autonomy of their will. Korsgaard (2009) indeed argues that humans have a form of self-consciousness that gives us a 'capacity to control and direct our belief and actions, [...] and makes us active in a way that [other animals] are not' (p.xi). As rational agents, we 'are faced with the task of making something of [ourselves]' (p.xii). Within this perspective, if an individual is characterised by coherent preferences, then preference satisfaction matters to the extent that it allows the individual to obtain what he *decided* he wanted to obtain. We argue in this section that acknowledging the faculty for the individual of 'being his own master' (Kant, 1797, p.30) implies that BWE offers a wrong diagnosis of the normative issue faced by boundedly rational individuals: the normative issue is indeed not that individuals are not able to maximise their welfare (either objectively measured in terms of happiness, or subjectively assessed as a total subjective comparative evaluation) but that they may lack of autonomy.

Recall that BWE adopts a third-person perspective so as to provide normative assessments. Adopting such a neutral perspective — coined by Nagel (1986) as the 'view from nowhere' — implies that the external third-person spectator must have at her disposal an external criterion for her normative evaluation (Carrasco, 2011), such as the quantity of happiness or the satisfaction of individual preferences. This is for instance on this basis that Conly (2013) argues that autonomy 'has been overvalued' (p.16), defining autonomy in the sense of Feinberg (1986), as 'the right to make choices and decisions — what to put in my body, what contacts with my body to permit, where and how to move my body though public space, how to use my chattels and personal property, what personal information to disclose to others, what information to conceal, and more' (p.54). Conly's central claim is indeed that letting people make their own choices may lead them to cause harm to themselves: there is therefore an *agent-neutral reason*[4] that justifies paternalism, the claim that satisfying the true preferences of any individual is normatively desirable. The satisfaction of one's true preferences is arbitrarily defined as the normative criterion for the third-person spectator, and therefore provides a reason for not respecting individual autonomy. Individual autonomy is however only

[4]A reason is defined as *agent-neutral* (by opposition to an *agent-relative* reason) if it does not derive from a normative fact concerning an individual in particular. Nudging young-Oscar on the basis that it will maximise happiness is an agent-neutral reason: promoting welfare (even subjective welfare) is indeed a reason that can be stated for another individual than Oscar. But nudging young-Oscar on the basis that it will induce him to quit smoking (because he will not be able financially to continue buying cigarettes) is an agent-relative reason: the formulation of the reason indeed essentially refers to Oscar himself and his addiction. On the differences between agent-neutral and agent-relative reasons, see Parfit (1984) and Nagel (1986).

considered in an instrumental perspective, as providing welfare *per se* (I may indeed prefer making my choices on my own), or as a necessary condition for one's self-development (which is beneficial *in fine* for all the society, since it for instance does not bound individual originality, which is necessary for society to progress). Questioning the validity of autonomy as a means to satisfy one's true preferences however does not imply that a claim for autonomy is not normatively grounded: it is indeed possible to justify the claim for autonomy from a 'second-person standpoint' (Darwall, 2006a), defined as 'the perspective you and I take up when we make and acknowledge claims on one another's conduct and will' (p.3). The second-person standpoint requires being able to put oneself in another's shoes so as to simulate the reasoning of others and attribute them mental states (see e.g. Goldman (1989, 1992), Gordon (1986, 1992)). Acknowledging the possibility for the individuals to simulate the reasoning of others implies in particular that we do not need to define an external normative criterion, as with the third-person perspective of standard normative economics[5]. Darwall (2006a,b) sees for instance in the second-person standpoint the roots of our moral responsibility, and then of the normativity of individual autonomy as part of respect for the dignity of persons: the autonomy of the will gives to the agents the ability to endorse a second-person standpoint, and then 'the authority, as a person, to make claims and demands of one another as rational and free' (Darwall, 2006b, fn.11). They have therefore the right to claim to be allowed to make their own choices: autonomy is therefore valuable for itself, and not valuable in an instrumental perspective — as suggested by Conly (2013) — as a means to promote welfare. The normative issue of paternalism, even if one's true preferences are not satisfied, is therefore that it is 'a failure of respect, a failure to recognize the authority that persons have to demand, within certain limits, that they be allowed to make their own choices for themselves' (Darwall (2006b, p.268), see also Shiffrin (2000) for a similar argument).

The normative criterion we suggest is therefore not the maximization of one's self-assessed well-being, but the development of *individual autonomy*. Before going further, and so as to provide a definition of 'autonomy', let us clarify our conception of individual agency — and how our approach differs from the model of the inner rational agent. We derive our model of agency from Dietrich and List (2013a,b,c,

[5]Darwall (2006a) for instance argues that Smith (1759) notion of empathy makes him 'one of first philosophers of the 'second person', if not the very first' (p.46): this leads Carrasco (2011) to argue that Smith's impartial spectator is not in a third but a second-person perspective, and that 'when judging from inside the situation, the spectator perceives some qualities that are unreachable to the external observer, and one of them is, precisely, propriety'.

2014) analysis of preferences and beliefs in rational choice, according to which individual preferences depend on certain 'motivationally salient' properties of the alternatives under consideration. The agent does not have any preferences *a priori*, and builds her preferences according to her perception of the choice problem. The agent then focuses (consciously or not) on certain motivationally salient properties of the alternatives. Those properties provide to the individual motivating reasons for preferring an alternative to another. The agent is then characterised by a weighing relation over property combinations, indicating how she ranks different property combinations. Preferences may then be incoherent since the salience of the properties is likely to change over time. Consider for instance Joe's choice in Sunstein and Thaler cafeteria. The items 'fruit' and 'cake' are characterised by several properties such as their position (one of the item may be more prominently displayed), their taste, their amount of calories, etc. When entering the cafeteria, only some properties of the items are salient from Joe's perspective. Suppose that the motivationally salient properties are the taste (according to which the cake is preferred) and the amount of calories (the fruit is preferred). Joe must then weight the different reasons for his choice, and chooses in accordance with this weighing relation. The ranking of the alternatives resulting from Joe's weighing relation are defined as Joe's preferences.
According to rational choice theory, Joe is aware of all the properties of the items, and attributes subjective weights to all those properties (his preferences are therefore a total comparative subjective evaluation). It is assumed that his preferences are consistent and context-independent: there are therefore some properties of the items that are assumed to be irrelevant (whose weights are null), such as the relative position of the items on the counter. Behavioural findings however highlight that Joe is not necessarily aware of all the properties of the items, and that he is likely to put non null weights (consciously or not) on what SuperReasoner would have considered as irrelevant properties. BWE then tries to reconstruct SuperReasoner's preferences, by assuming that he has access to a mode of reasoning that will allow him to define a unique weighing relation, with null weights on the properties related to the choice architecture.

We suggest that the normative issue raised by behavioural findings is that Joe's choice may be influenced by salient properties he is not aware of. He can therefore indirectly be manipulated by the choice architect, *via* the manipulation of the choice architecture. We therefore define Joe as an *autonomous agent* if Joe is *aware* of the reasons that drive his choices, and if he *accepts* that those specific reasons drive his choices. If Oscar chooses the default option simply because it is the default

option (without knowing that default options are more salient, independently of their content), then he is not autonomous: his final choice is indeed shaped by the choice architect, without his consent. But if Oscar is perfectly aware that he is inclined to choose the default option simply because it is more salient, he is also able to choose to ignore this property (and then to focus on other motivationally salient properties). Oscar is autonomous here, because he chose to accept that his choice could be shaped by the choice architect.

Note that the idea that people can 'accept' their own preferences can be understood in two different senses, according to whether we are referring to a Humean or a Kantian account of reasoning. In the first case, my reason is the simple spectator of my passions, i.e. of the 'reasons' that drive my choices: this notion of acceptance simply means that the individual accepts to live with her passions, but also that she knows that she has absolutely no control on those passions. In the second case, I can actively choose the reasons that drive my choices: the notion of acceptance is therefore stronger, since it also entails the possibility to choose one's own preferences. Although our argument will share many features of the Kantian position (in particular the idea that people can choose their own preferences, and that the normative value of individual autonomy is second-personally grounded), our picture of the autonomous agent will be significantly different from Hausman and Welch's inner rational agent.

Unlike the multiple selves model of BWE (distinguishing between a *Homo economicus* and a *Homo psychologicus*), Dietrich and List's model can be seen as using the concept of *metaranking* (Sen, 1977): the individual does not always act on the same preference ordering (because the salient properties are not always the same), and the person is unified by the existence of a coherent higher-level ranking of the lower-level rankings (the weighing relation over the different combinations of properties). However, instead of considering that this higher-level ranking is given, we assume that this ranking is *chosen* by the agent: the person is therefore unified by the continuity of agency of his successive selves, who chooses the life that, as an agent, she wants to live[6].

Our notion of autonomy is grounded on two different ideas: (i) the individual is aware of the possible factors that may interfere with her deliberation, and (ii) the

[6]We present in chapter 5 the sketch of a general theory of preference formation that would not require the existence of metapreferences to determine the choice of one's own preferences. The basic idea is that the individual can generally benefit from strategic commitments (i.e. from acting according to preferences that are not her material payoff): what I perceive at a certain date as my self-interest determines my 'strategic' preferences, which then progressively becomes my material payoff, allowing for a new strategic commitment and the formation of new preferences.

individual can actively choose the reasons that should matter for her choices. Note that there exist situations in which only the first condition is verified, typically in cases of addiction. Suppose that Oscar knows that the main reason that prevents him from quitting smoking is his addiction to nicotine. Although he is perfectly aware of it, he does not accept that his behaviour is driven by his addiction. Oscar is therefore not autonomous in this situation, since his behaviour is shaped by a reason he does not accept.

Our definition of individual autonomy is a *responsiveness-to-reasons* account of autonomous agency (e.g. Wolf (1990), Fischer and Ravizza (1998), Nelkin (2007)), according to which 'an agent does not really govern herself unless her motives, or the mental processes that produce them, are responsive to a sufficiently wide range of reasons for and against behaving as she does' (Buss, 2014). Furthermore, when she is able to choose her own weightiest reasons, and therefore to see herself as an agent, an individual responsive to reasons is also morally responsible for her actions (Fischer and Ravizza, 1998). Nudges are therefore unacceptable, because they induce behavioural changes that we are not aware of, and therefore for which we cannot take responsibility.

It is worth noticing that, unlike Hausman and Welch (2010), we do not need to refer to the inner rational agent to define our notion of autonomy. Indeed, the agent we are considering (whose autonomy can be bounded) is not a counterfactual agent defined as what the individual would be if she were perfectly rational: the agent is the actual individual, aware that her behaviour is driven by many psychological processes, on which she has a limited control, but who nevertheless accepts to live with this fact of human psychology. The main distinction between those two approaches is that HW notion of autonomy is a *responsiveness-to-reasoning* account of autonomous agency (Christman, 1991, 1993, Mele, 1993, 1995): an agent is autonomous if she is able to critically evaluate her motives and reasons on the basis of her beliefs and desires, and to adjust them in the light of her evaluation. The autonomous agent would therefore be in the position of the inner rational agent: this implies that her preferences are necessarily internally consistent (properties such as completeness and transitivity are indeed a logical consequence of comparative judgements (Hausman, 2012)). On the contrary, we argue that an autonomous agent can hold incoherent preferences, if she knows why her preferences are incoherent, and take responsibility for her preferences.

As behavioural economists, we know that people may act on reasons they are not aware of (such as trying to avoid losses at any cost). We cannot however be sure that people are not aware of those phenomena, and also that, if they were

aware of them, they would change their behaviour (they have indeed the right, as autonomous agents, to accept holding incoherent preferences). Instead of helping the individuals by nudging them towards what we think they would prefer if they were rational, we suggest helping the individuals to become autonomous, and then let them make their own choices. This means simply that individuals should be informed about the possible psychological processes that are likely to interfere with their deliberation, such as framing effects, and that they should have the possibility to choose the reasons that determine *in fine* their choices. Those two dimensions can be understood as improving individual freedom in a negative and a positive sense (Berlin, 1958)): the choice architect objective should be to develop individuals' critical thinking (such that they will not be influenced by factors they are not aware of), and also to ensure that the individuals have the opportunity to choose one of the lives they may have reason to value. Our recommendations are therefore in line with education policies whose core goal is the development of individual autonomy (e.g. White (1982), Gutmann (1987), Kamii (1991), Cuypers and Ishtiyague (2008)).

Consider for instance the case of a doctor who must describe different possible treatments to a patient: it is likely that the patient does not consider the presentation of the treatment in terms of probability of success or failure as a weighty reason for her choice. Rather than choosing the choice architecture that — according to her — will improve the well-being of the patient, the doctor could for instance present a single information with different frames (for instance present the probability of success of each treatment and then the probability of failure, while emphasising that the final choice can be influenced by the way with which the information was displayed). The patient would become aware of the existence of framing effects, and is then likely to take a better decision, without being misled by features that she can decide to consider to be irrelevant for her choice[7].
Consider now the cafeteria of Sunstein and Thaler. Rather than directly choosing the location of the different items, Carolyn could inform the users of how their choices can be affected by the choice architecture, and then give them the opportunity to choose the choice architecture (by a public discussion and a vote for instance). If the users show little interest in this question and do not want to get involved in the choice of the location of the different items, they can delegate

[7]We can highlight the similarity between this last situation and Luke's investment choice discussed in chapter 3. Rather than choosing a choice architecture that is likely to influence Luke's choice, Claire should inform Luke of the potential manipulative power of framing effects, and tell him that a choice between options (A) and (B) is equivalent to a choice between options (C) and (D).

their choice to Carolyn and let her in charge of designing the choice architecture. This would mean that, as SuperReasoner, they have no weightier reason to prefer the fruit (or cake) to the cake (fruit): their choice are therefore guided by more important reasons, such as being able to choose rapidly one's dessert (a fruit or a cake). The main difference between Sunstein and Thaler's example and this last situation is that, although Carolyn manipulates *in fine* the choice of the individuals in both cases, she has their consent only in the second case, i.e. when the individuals had the opportunity to choose the location of the items but preferred to let a third party in charge of it.

Quite interestingly, it appears that Sunstein and Thaler anticipate the criticism presented here by arguing that '[the most ardent libertarians] are concerned about liberty and free choice rather than about welfare' and therefore that 'they prefer required choosing to nudges' (Thaler and Sunstein, 2008, pp.242-243). Their argument is then that it is not necessarily wise to force people to choose, since 'people would often choose not to choose' in particular when 'the choices are hard and the options numerous' (p.243). Our criterion of individual autonomy is however slightly different from what they perceive as a libertarian objection, since we argue that people should be able to choose whether they want to choose or not (they are indeed still autonomous if their behaviour is shaped by reasons they accept, and can therefore choose not to choose). The normative issue of LP is that it forces people not to choose whether to choose or not (the choice architecture is indeed taken in charge by the choice architect, and the individual has no power on this decision — although she is the one who will be *in fine* influenced by the choice architecture).

## 4.4 Normative economics and democracy

Focusing on the development of individual autonomy rather than preference satisfaction questions the standard view that normative economics is addressed to a social planner or choice architect (whose objective is to ensure the satisfaction of individual preferences), rather than directly to the individuals (McQuillin and Sugden, 2012b, p.556). We would like to argue now that normative recommendations should be conceived as advices to the citizens rather than policy recommendations addressed to an abstract social planner.

### 4.4.1 The Social Planner and the Leviathan

By endorsing the social planner position, economists simultaneously endorse two roles: defining what is socially desirable, and how to achieve a socially desirable state. The Sovereign, as the sharer of the supreme authority over society, is the only entity legitimate to define what matters for the society, while the Government, as the depositary of the Sovereign's authority to implement her will, is in charge of designing the society to achieve the Sovereign's will. Taking the viewpoint of the social planner implies that economists define the normative criterion of the society: but are they legitimate to replace the Sovereign in her normative assessments? We would like to argue here that this position is justifiable within the context of the hive, when individuals have exogenous and coherent preferences, and is consistent with Hobbes contractarianism. However, as soon as we consider that individuals have some autonomy and can choose to some extent their own preferences, we should abandon this perspective: this will enable us to pass from a Hobbesian analysis of the social contract and of the role of the social planner to Rousseau's contractualism and the central place of democratic processes in the definition of individual preferences.

Suppose that the world can reasonably be represented by the hive described above, and that the individuals — as in conventional welfare economics — are actuated by the satisfaction of exogenously given preferences. Although the first fundamental theorem of welfare economics ensures that the free interaction of the individuals leads to a socially desirable outcome in competitive markets, they are not able to reach a Pareto efficient outcome if they face collective action situations or interact in an imperfect market. This suboptimal situation is quite similar to Hobbes' *state of nature*, and his idea of a 'war of all against all': there does not exist any entity at the collective level to ensure the pacific coexistence of all the individuals, whose interests are generally conflicting (they have in particular no reason to trust each other, since they cannot be sure that the others will respect their promises). The hive is therefore in a suboptimal state of anarchy, and it would be in the interest of each individual to define an entity at the aggregate level such that the satisfaction of its collective preferences would lead to a Pareto-superior outcome. Such an entity would be the *Sovereign* of the society, since its preferences would define what is preferred by the society as a whole. But how can we define the collective preferences of the Sovereign? Or put differently, how can we define what matters at the collective level, given the preferences of each individual (whose satisfaction matters)? The difficulty of this question is that Arrow's impossibility

theorem precisely states that such collective preferences cannot exist if we want to verify some basic and intuitive axioms: non-dictatorship, universality, independence of irrelevant alternatives and unanimity (Arrow, 1951).
The first solution suggested so as to 'save' the hive from this suboptimal state of anarchy is the advent of a dictator. This is for instance Hobbes' position, who considers that men are fundamentally unable to coexist peacefully without a supreme authority. Hobbes conceptualised the social contract in those terms:

> This is more than consent, or concord; it is a real unity of them all in one and the same person, made by covenant of every man with every man, in such manner as if every man should say to every man: I authorise and give up my right of governing myself to this man, or to this assembly of men, on this condition; that thou give up, thy right to him, and authorise all his actions in like manner (Hobbes, 1651, p.106)

The individuals therefore voluntarily give up their 'right to govern [themselves]' to another individual or group of individuals, on the condition that everyone does the same. In this situation, the collective preferences are defined exclusively by the preferences of a single individual (or of a subgroup of the population). The different individuals must then respect the will of the Sovereign: although the Sovereign's preferences may not be consistent with their own preferences, this situation is better than the state of anarchy without Sovereign (they achieve here a second best). This may be why the individuals merely *authorise* the Sovereign to govern them: the power of the Sovereign over the rest of the individuals is not absolute, since it must provide them more than what they would have gotten in the state of nature (safety and peace according to Hobbes). In a standard microeconomic context, if the market equilibrium is not Pareto efficient, the individuals authorise the social planner to implement public policies such that the new equilibrium is a Pareto improvement. They have however nothing to say about the way the social planner shares the surplus between the individuals: the collective preferences of the social planner are indeed arbitrarily defined by the modeller, in general as a sum of individual utilities.

An alternative solution is grounded on the observation that in competitive markets, the initial situation of anarchy within the Hive is actually not suboptimal: it seems therefore possible to ensure the peaceful coexistence of the individuals, and this without coercing their freedom by designating a dictator. It is likely that Hobbes did not think of this solution because he believed that the interaction of self-interested individuals could only lead to an open conflict: Gauthier (1969)

indeed argues that 'it is impossible to emphasize too strongly that it is the substantive premises about human nature, and not the formal structure of the theory, that determines its absolutist character'. If self-interested individuals can coexist and reach a desirable state in competitive markets, then a solution to improve the society while avoiding the appeal to a dictator is to build markets, i.e. to ensure that all the interactions in society are ruled by competitive markets. This is precisely the approach defended by neoclassical economics.
However, just as Gauthier argues that the absolutist character of Hobbes theory is due to its anthropological premises, we can argue that this reasoning is valid if and only if the underlying anthropological model of the Hive makes sense. In particular, humans should be passive actors fundamentally guided by their self-interest. The reason why social cooperation – although it is clearly in the interest of each individual – cannot be naturally reached is that the individuals described in the hive are not able to form commitments. They cannot decide to go beyond their immediate interest, since they are programmed to systematically choose the action that satisfies their best interest. The Leviathan – through its absolutist character – confers them indirectly this power of commitment[8]. Alternatively, competitive markets can replace the social dimension of exchange (that could generate some conflict) with a disembodied price system: a perfectly competitive market would be a 'morally free zone, a zone in which the constraints of morality would have no place' (Gauthier, 1986, p.84). Indeed, according to Gauthier, since the common good is an unintended by-product of individuals' pursuit of their self-interest in competitive markets, market relationships do not need to be genuinely social (Bruni and Sugden, 2008, p.38).

Assuming that individual preferences are exogenously given seems therefore to legitimate the standard approach in normative economics, according to which we should take the standpoint of a benevolent despot so as to provide normative assessments. The normative criterion may not be in the best interest of each player (such as for instance the maximisation of the sum of utilities, since each player would individually prefer to have a higher weight than the others), but they should however be ensured to get at least the level of utility they would have achieved at the market equilibrium without intervention. Neoclassical economists can therefore

[8]Hobbes's position is that cooperation among large groups is impossible in the absence of a supreme political authority, since the threat of sanctions is necessary to give the insurance to everyone that the others will cooperate: it is not because not cooperating is too costly that cooperation is ensured, but because it is actually rational for each individual to cooperate when they are ensured that all the others are committed to reciprocate (Hollis, 1998).

defend their approach on the basis that it provides a contractarian solution to the initial suboptimal equilibrium: the existence of the social planner (and therefore the legitimacy for economists to endorse this position) is justified because it enables the individuals to reach a mutually beneficial outcome, compared to the initial situation without intervention. This approach however does not respect individuals' autonomy and their right to decide by themselves the terms of the agreement (we can indeed notice that the choice of the normative criterion is left to the economist's discretion, without any control of the individuals).

### 4.4.2 Autonomy and the social contract

Suppose now that, unlike within the model of the hive, individual preferences are not necessarily fixed, and players can make commitments. We can now provide two other solutions to improve the initial equilibrium. The first solution is that rational individuals may decide to make the commitment of not breaking agreements, by grounding principles of morality on contractarianism (Gauthier, 1986, 1991). Unlike Hobbes who suggested that the existence of the Leviathan was required to provide the insurance to all the individuals that the other individuals would be committed to cooperate, Gauthier argues that rational individuals have the ability to choose to follow principles of morality that go against their direct self-interest, but that allow them to reach mutually beneficial agreements with other 'moral' players. Gauthier indeed suggests that rational players may either be *straightforward maximisers*, who always choose what is in their best interest (and may therefore break non binding agreement such as 'cooperate in a prisoner's dilemma when the other cooperates'), or *constrained maximisers*, who make the commitment of not breaking agreements. Gauthier then argues that, in a world in which the type of maximiser (constrained or straightforward) of each individual is perfectly observed, then constrained maximisers will outperform straightforward maximisers[9]. In this situation, the individuals do not need a benevolent despot to enforce their commitments: it is indeed in their own interest, as rational agents, to choose to become constrained maximisers, because it throws the foundations of reciprocity within the society.

The second solution is slightly different from Hobbes and Gauthier's contractarianism, since it offers a contractualist solution to the initial suboptimal equilibrium. The distinction between contractarianism and contractualism is that

[9]This result is formally shown in chapter 7, proposition 9.

contractarianism takes moral principles to result from rationally self-interested bargaining, while contractualism sees the agreement as governed by a moral ideal of equal respect (Darwall, 2003, p.4). Contractarians indeed assume that they have a natural right to claim the outcome they would achieve in the absence of agreement: the agreement is then build as a mutually beneficial agreement on the basis of this disagreement outcome. Contractualists, on the other hand, consider that this moral claim on the disagreement outcome is arbitrary, and therefore that the moral principles resulting from the bargaining on this initial outcome have no moral force. Contractualists are not trying to gain acceptance from rules they prescribe from the perspective of their own interest, but they are prescribing and agreeing on rules from a common perspective as one free and equal person among others. This is typically Rousseau's position and his conception of the political community of citizens, as a form of association in which each, 'uniting with all, nevertheless obeys only himself'. The Kantian categorical imperative rests on a similar logic, since the moral law the individual prescribes to herself should become a universal law. A similar approach is developed by Rawls (1971), since the original position, under the veil of ignorance, implies that the individuals are disconnected from their personal interests (although they are deliberating so as to satisfy their individual interest, they have no idea of what will be their position in the society). The principles of justice they are choosing are therefore chosen among a community of free and equal individuals. More recently, Parfit (2010) suggests the *Kantian Contractualist Formula* as an explicitly contractualist reformulation of Kant's categorical imperative, according to which 'everyone ought to follow the principles whose universal acceptance everyone could rationally will' (p.20).

An interesting difference between the contractarian and the contractualist approaches is that, while the notion of *authorisation* is central with Hobbes, contractualists like Rousseau put a strong emphasis on the notion of *commitment*:

> Each of us puts his person and all his power in common under the supreme direction of the general will, and, in our corporate capacity, we receive each member as an indivisible part of the whole. [...] This formula shows us that the act of association comprises a mutual undertaking between the public and the individuals, and that each individual, in making a contract, as we may say, with himself, is bound in a double capacity; as a member of the Sovereign he is bound to the individuals, and as a member of the State to the Sovereign. (Rousseau, 1762)

Rousseau's social contract gives rise to an emerging entity, the *People*, who is sovereign. Contrary to Hobbes who designated the Sovereign among the population — i.e. designated a dictator — Rousseau considers an emerging entity, the group as a distinct agent. The People has his own will, the *general will*, which can only be revealed by the enlightened deliberation of the citizens. Unlike the contractarian approaches in which the object of the contract can be seen as the fair aggregation of individual preferences, so as to reach a mutually beneficial outcome, the social contract consists in a transfer of agency from the individual to the collective level through a collective act of commitment (Hollis, 1998): the individuals can then choose to become an 'indivisible part of the whole', so as to be actuated by the collective objective of the group rather than by their own interest.
Although the contractarian and the contractualist approaches differ in their motivations (contractarians are indeed ultimately self-interested, while contractualists are interested in their community), they both induce a transfer of agency from the individual to the collective level: the individuals are looking for a *mutual advantage*, instead of trying to satisfy their own preferences. Social interactions, unlike within the hive where the only motive of action is the satisfaction of one's own preferences, can be understood as joint intentions for mutual assistance (Bruni and Sugden, 2008). Individuals looking for mutual advantage rather than their own self-interest can then engage in *team reasoning* (Sugden, 1993, Bacharach, 2006), i.e. choose their action by considering themselves as a part of a larger entity, the group of individuals within which the agreement is recognised. Although team reasoning is an end in itself for a contractualist, it is only instrumental for the contractarian, as a means to achieve *in fine* her own interest.

On a more pragmatical level, it is not sure whether the individuals are fully aware of the objective of the others: they must then discuss to identify what outcome could be mutually beneficial (this is not necessary within the hive, since the individuals do not care about the objective of others — they are only interested in the satisfaction of their own preferences). It is only after this phase of identification of the interest of each individual that a consensus may emerge concerning the possible mutually beneficial outcomes. The viewpoint from which normative assessments are made is therefore not a 'view from nowhere', outside society, but a 'view from everywhere', everywhere within the society. Reasons for actions are then agent-relative rather than agent-neutral: normative claims are indeed directly related to the actual agents of the group, and not to a normative principle stated by a third-person observer.
As discussed above, behavioural findings highlight that actual individuals may not

be able to have an enlightened judgement, and therefore to successfully participate in public debates: so as to ensure the continuity of the People, the government should take in charge the political education of the citizens, to '[prepare] citizens to participate in consciously reproducing their society' (Gutmann, 1987, p.287). Since individuals are likely to lack of autonomy, to be influenced by reasons they do not recognise as being their owns, and since that, as citizens, they are the only source of normative authority in society, normative economists should directly advise the citizens, not only about how to pursue what they consider as their common interests, but also about how to deal with the diverse biases that are likely to interfere with their deliberations.

This alternative perspective on normative issues also gives us a way to precisely define the cases in which individuals make mistakes, such as not looking to one's right before crossing a street in London. The reason why this situation can be considered as a mistake is not because there exists an external observer able to determine objectively what a mistake is, but because, as an autonomous agent, I can endorse the standpoint of other autonomous individuals (with their own subjective and not necessarily coherent preferences), and reasonably consider that, from the perspective of *any* autonomous individual, not looking to one's right is a mistake. On the contrary, although I may consider that hyperbolic discounting is a mistake, I can also endorse the standpoint of an autonomous individual (e.g. Derek Parfit) for whom hyperbolic discounting is not a mistake. A specific action is therefore a mistake if and only if any autonomous individual (able to simulate the reasoning of other autonomous individuals) would consider this action as a mistake. Instead of unilaterally deciding that saving little for one's retirement is a mistake, we should simply engage in a discussion with other individuals so as to confront our personal judgements. This suggests that, for all the situations in which economists can reasonably assume that any autonomous individual would consider a specific choice as a mistake, then a nudge may be justified. However, there currently exist many cases such as savings choices, for which many autonomous individuals do not share the diagnosis of behavioural welfare economists: in those cases, an intervention cannot be justified as a case of means paternalism.

Neoclassical and behavioural welfare economists, by trying to assess society from an outside position, may *in fine* impose their own normative views to other individuals — they are indeed not able to view the society from this third-person perspective. The only solution to produce normative assessments is to confront personal views within the society: normative economists, rather than trying to guess what matters for the individuals, should try to organise society such that the citizens can deliber-

ate, and then construct their own normative claims. Economists are not legitimate to decide what matters for the individuals: they should instead help the citizens to debate so as to identify mutually beneficial agreements. Their role is therefore not to replace the Sovereign and to advise the Government, but to directly advise the Sovereign on the mutually beneficial outcomes that may be reached.
The distinction we draw between neoclassical welfare economics – grounded on a social planner assessment – and our procedural form of normative economics is perfectly illustrated with the dual interpretation we can have of the capability approach (Qizilbash, 2011, Baujard and Gilardone, 2015). While Sen (2009) emphasises that what fundamentally matters is public debate and democracy, other authors – such as Nussbaum (2000) – place themselves in the position of an expert and want to define a universal list of capabilities that could characterise a 'good life'. The latter – considered today as 'the' capability approach – is therefore endorsing a social planner perspective, aiming at defining what matters for the individuals, whereas the former considers that providing basic capabilities is not an objective in itself, but a necessary condition for the individuals to be able to make public debates.

### 4.4.3 The management of common-pool resources

We now illustrate our claim by considering, on a more empirical level, the design principles that characterize robust institutions for managing common pool resources (Ostrom, 1990). We want to highlight that the sustainable management of CPR is possible if and only if public policies are designed such that they provide to the individuals the means to organise themselves, rather than imposing from the outside a solution designed for correcting a market failure by the introduction of adequate incentives.
CPR are a class of goods characterized by two attributes, the difficulty of excluding individuals from benefiting from the resource, and the subtractability of the benefits consumed by an individual from those available to others. Two main types of problems can emerge in this context, appropriation and provision problems: appropriation problems are related to the exclusion of potential beneficiaries and the repartition of the output, whereas provision problems are related to the management of the stock of the resource, whether it be its creation, the maintenance or improvement of its production capabilities, or the avoidance of its destruction (Ostrom et al., 1994, p.9). Ostrom (1990) suggested a list of eight design principles that characterize the institutions enabling a sustainable management of CPR, which have been slightly amended by Cox et al. (2010), who provide a meta-analysis of the different empirical works that tested those principles (extract from Cox et al. (2010,

table 4)):

- 1A, user boundaries: clear boundaries between legitimate users and non users must be clearly defined;
- 1B, resource boundaries: clear boundaries are present that define a resource system and separate it from the larger biophysical environment;
- 2A, congruence with local conditions: appropriation and provision rules are congruent with local social and environmental conditions;
- 2B, appropriation and provision: the benefits obtained by users from a CPR, as determined by appropriation rules, are proportional to the amount of inputs required in the form of labour, material, or money, as determined by the provision rules;
- 3, collective-choice arrangements: most individuals affected by the operational rules can participate in modifying the operational rules;
- 4A, monitoring users: monitors who are accountable to the users monitor the appropriation and provision levels of the users;
- 4B, monitoring the resource: monitors who are accountable to the users monitor the conditions of the resource;
- 5, graduated sanctions: appropriators who violate operational rules are likely to be assessed graduated sanctions (depending on the seriousness and the context of the offense) by other appropriators, by officials accountable to the appropriators, or by both;
- 6, conflict-resolution mechanisms: appropriators and their officials have rapid access to low-cost local arenas to resolve conflicts among appropriators or between appropriators and officials;
- 7, minimal recognition of rights to organize: the rights of appropriators to devise their own institutions are not challenged by external governmental authorities;
- 8, nested enterprises: appropriation, provision, monitoring, enforcement, conflict resolution, and governance activities are organized in multiple layers of nested enterprises

Our purpose is not to discuss extensively these different principles, but to highlight that most of them are directly supported by our conception of normative economics as developing democratic processes rather than implementing incentives. We can indeed notice that the main feature of those principles is the idea that the users of the CPR should be able to design their own institutional environment (this is quite explicit in the principles 3 and 7). Furthermore, the possible external actors who monitor the users and the resource, or who assess possible sanctions in case of non-respect of the appropriation and provision rules are systematically accountable to the users (who therefore remain sovereign). Several empirical studies showed for instance that when the rules are imposed by an external authority, this one generally fails to enforce them, leading to suboptimal results (Ostrom et al., 1994, pp.221-222). Nevertheless, although direct interventions often fail, the government can help the users to manage more efficiently the resource: Blomquist (1994) – from empirical evidence of groundwater systems in Southern California – suggests for instance that the design of provision and appropriation rules is facilitated by the presence of government agencies that can provide reliable information to the users (pp.296-297). From various laboratory experiments and field studies, Ostrom et al. (1994) argue that the individuals can overcome the temptation of overusing the resource if they have some expectation of mutual trust, or the possibility of building trust through continued interaction and communication (p.328), and if they have some autonomy to decide on their own rules (p.323). However, since it appears that boundedly rational individuals can have some difficulties to reach optimal rules – mainly due to information issues and the complexity of the problem – governmental agencies play an important role by recognizing the right to the individuals to form their own rules and commitments, but also by providing them reliable information and backup enforcement mechanisms (pp.322-327).

We can now notice that those conditions, and in particular the role of the government as an actor who provides information and support to the individuals without directly intervening nor trying to influence individuals' choices, correspond to the kind of normative prescriptions that would result from our conception of normative economics. Our claim is indeed that economists should assess public policies in terms of individual autonomy, i.e. the ability of the individuals to engage in public debate as informed citizens, so as to be able to collectively choose their own preferences: this requires providing the largest information to the individuals, and let them decide on their own rules rather than imposing external rules.
In addition, the case of CPR gives us another argument in favour of a more deontological formulation of normative economics, the impact of institutional rules

on individual preferences. It seems indeed that individual preferences in CPR situations depend on the institutional organisation that rules the appropriation and the provision of the resource: self-organized institutions are more likely to generate prosocial behaviours than rules imposed by an external authority. It means that imposing the same policy can have a different impact according to its initiator: empirical evidence in CPR situations suggest that policies democratically implemented are more likely to be efficient than policies implemented by an external authority. It is therefore probably not equivalent to try to implement what the individuals would have chosen if they were autonomous (such as within a social planner's perspective) and to try to directly improve the autonomy of the individuals. A measure implemented by the individuals who will be directly affected by it may be more efficient than the same policy implemented by an external authority: in the latter case, the individuals can indeed be suspicious about the objective of the government, and then be affected by crowding-out effects (they may for instance be tempted to cheat and exploit the weaknesses of a system of monetary incentives implemented by an external authority).

The management of CPR offers us a good illustration of one of the main objectives of our reformulation of normative economics. While neoclassical welfare economists ground their normative assessments on consequentialist considerations such as the welfare generated by the satisfaction of one's preferences, we suggest adopting a more procedural approach by grounding our normative assessments on individual autonomy and the ability for the individuals to choose themselves what matters for them. It seems indeed that the sustainable management of a CPR (and therefore the welfare it generates) is not only the result of the implementation of specific rules, but also of the conditions under which those rules were decided: promoting individual welfare therefore requires promoting individual autonomy, since the rules that will enable the individuals to maximise their welfare are more likely to be efficient if they are implemented by autonomous agents rather than by an external authority.

# Part II

# A Model of Endogenous Preferences

Chapter 5

# Choosing One's Preferences

**Contents**




**Abstract:** we develop in this chapter a model of preference formation, that will constitute the basis of the general model of preferences we will present in chapter 6. We assume that the players are able to choose in a precommitment game the weights they assign to the other players' material payoff in their own preferences, and determine the optimal weights each player should choose so as to maximise her material payoff. We highlight a systematic relation between supermodularity (submodularity) and the formation of cooperative (competitive) preferences. We then investigate the possible implications of this framework for the design of public policies. We show in the case of climate change negotiations that international agreements relying on technology standards with trade sanctions rather than objectives

of pollution abatement are more likely to succeed, since they create a coordination game and cut the strategic substituability of the initial game — that would have given incentives to adopt more competitive preferences.

## 5.1 Introduction

By considering preference inconsistencies and deviations from the axioms of rational choice as mistakes, BWE makes the implicit claim that individuals would be better off if they were behaving as prescribed by rational choice theory. The aim of this chapter is to show that this claim is generally false in the case of strategic interactions. We indeed show that players are generally better off when acting 'irrationally' than when following the recommendations of rational choice theory. The key mechanism supporting this result is that rational choice theory can be indirectly self-defeating (Parfit, 1984). There indeed exist games in which a person who is perfectly rational, trying to maximize her utility, can achieve *in fine* a lower payoff that a less rational individual. The counter-intuitive implication of this phenomenon is that, in those specific games, if an individual wants to achieve her objective, then it is in her interest to adopt an apparently non rational behaviour: being irrational can therefore be rational in those games. So as to illustrate this point, imagine the following situation:

> **Symmetric Cournot game**: we are playing a symmetric Cournot game. Suppose that I decide to maximise my profit minus $\sigma$% of your profit (I want therefore to maximise my profit as well as the difference between our profits): you know that I am an aggressive player, and therefore that I am likely to produce more than at Nash equilibrium. You then reduce your production and we end up in a situation in which you play your best reply to my somewhat 'irrational' action. I am then producing more than my Nash output and you less: if I choose the adequate level of $\sigma$, then the resulting equilibrium can actually correspond to the Stackelberg equilibrium in which I would be the leader and you the follower.

In this situation, although I did not directly maximise my profit, we reached an outcome in which I obtained a higher profit than if I had directly maximised my profit. The idea that players can benefit from strategic commitments — i.e. voluntary deviations from the rational behaviour — have been suggested by Schelling (1960), and already discussed by von Stackelberg (1934), with the introduction of timing in oligopoly. Different approaches have then been developed

in order to study specific kinds of commitments, such as strategic delegation (e.g. Fershtman and Kalai (1997), Fershtman and Gneezy (2001), Sengul et al. (2012)), the evolution of preferences (e.g. Güth and Yaari (1992), Samuelson (2001), Heifetz et al. (2007a,b)), or the role of emotions in decision making (Franck, 1987, 1988). In addition, some experimental results suggest that players progressively learn to make the optimal strategic commitment (Poulsen and Roos, 2012).

From a methodological perspective, the common feature of those approaches is to provide a theory of the endogenous formation of individual preferences: they all rely on the more or less explicit existence of a 'preferences game', in which the players can choose their optimal preferences in a first stage, before playing the actual game with those distorted preferences. This idea is for instance explicit within the indirect evolutionary approach: it is indeed assumed that players consciously play the game with distorted preferences, and that a process of natural selection operates on the underlying preferences game (we will study more extensively the indirect evolutionary approach in chapter 7).

We develop in this chapter a model of endogenous preferences, that will constitute the basis of the general model of preferences we will present in chapter 6. In line with the works cited above, we will distinguish between one's *material payoff* and one's *strategic preferences*. While the material payoff measures the 'gain' of the individual for a given strategy profile, the strategic preferences correspond to the counterfactual preferences a rational player would hold so as to maximise her material payoff. For simplicity, we will assume here that the material payoff is an indicator of one's well-being — either objectively defined as the hedonistic experience of the individual, or subjectively defined as in BWE — and that the individual is ultimately actuated by the pursuit of her well-being. We will discuss in more details the interpretation of one's material payoff in the next chapter within the context of Bacharach (2003) variable frame theory. The point we intend to raise in this chapter is indeed a more technical one: we show that, even if we can define a coherent measure of one's well-being, it is not certain that preference inconsistencies are mistakes. We can indeed notice that BWE does not make any distinction between the notion of 'true preferences' understood as the measure of my subjective well-being (my material payoff), and understood as the preferences I would hold under TS's conditions (my strategic preferences). It is indeed implicitly assumed that my counterfactual preferences necessarily correspond to my material payoff, since a rational individual chooses her action so

as to maximise *in fine* her subjective well-being. The main point of this chapter is that, in the case of strategic interactions, there will generally exist a discrepancy between one's material payoff and one's strategic preferences. The notion of 'true preferences' is therefore ambiguous when studying strategic interactions, since it can either refer to the material payoff of the individual or to her strategic preferences.

We assume that individuals are able to choose their strategic preferences so as to better satisfy *in fine* their material payoff, and define the set of possible individual strategic preferences as the set of weighted sums of individual material payoffs. By determining the Nash equilibrium of this preferences game, we will obtain the optimal strategic preferences a rational player should choose so as to maximise her material payoff, knowing that the others may also make strategic commitments. Furthermore, we will also be able to identify the features of the strategic environment that could lead to the endogenous formation of cooperative and competitive preferences[1].
We firstly determine the set of games for which it is rational to choose null weights, i.e. for which the strategic preferences correspond to the material payoff: we show that, as soon as there exists a player $i \in N$ such that the Nash equilibrium is not the Stackelberg equilibrium of player $i$, there is at least one player who should choose non-null weights. We then determine the optimal weights, and show that players tend to develop cooperative preferences if and only if the game with their new preferences is supermodular. It means that players have a tendency to cooperate in presence of strategic complementarities and to compete against each other in presence of strategic substitutes. Since individual preferences will generally be different from one's material payoff, we discuss the implications of our findings for the design of public policies: we indeed argue that policies creating a coordination game between the players are more efficient than policies keeping the initial structure of the game (in case of submodular games, such as a public good game with a concave benefit function), since they lead to the endogenous formation of cooperative preferences. We illustrate those results by focusing on climate change negotiations.

The rest of this chapter is organised as follows. Section 5.2 develops our model of preference formation. Section 5.3 states our propositions, and section 5.4 investigates the impact of public policies on individual preferences, with an analysis of

[1]Unless a confusion is possible, we will simply use 'preferences' to refer to one's 'strategic preferences'

climate change negotiations. Section 5.5 concludes.

## 5.2 A model of endogenous preferences

In this section, we firstly clarify our interpretation of the notion of strategic preferences. We then introduce technical notations and define a notion of equilibrium characterising a strategy profile immune to individual strategic commitments. We illustrate those different notions by studying a public good game.

### 5.2.1 Bluff and commitment

We already presented above the logic of the two-stage game in a Cournot competition: since being aggressive with firm 2 may be *in fine* beneficial to firm 1, it is in the interest of firm 1 to choose its level of production not only to maximise its profit, but also to maximise the difference between its profit and the profit of its opponent. From the perspective of firm 1, its *relative* success may therefore matter more than its *absolute* success, since by trying to outperform firm 2 rather than merely maximising its payoff, firm 1 is likely to obtain a higher profit than if it has adopted a profit-maximising strategy.
A question that may arise is then the status we should give to those different types of preferences: if firm 1 decided to adopt strategic preferences different from its material payoff, can we still say that the true objective of the firm is payoff maximisation? It would actually be more accurate to say that the true objective of firm 1 is to beat the competition, and that a fortunate by-product of this objective is the maximisation of firm 1's profit: it is only because firm 1 is committed to be aggressive that its profit is maximised (otherwise firm 2 would anticipate that firm 1 is bluffing, and therefore that firm 1 will play *in fine* its best reply). A salient illustration of this difference between commitment and bluff can be found in the current negotiations between Greece and the European Union concerning Greek national debt (and more generally in bargaining games). In a Op-Ed article in the *New York Times*, Greek finance minister (and game theorist) Yanis Varoufakis (2015) claims that 'it would be pure folly to think of the current deliberations between Greece and our partners as a bargaining game to be won or lost via bluffs and tactical subterfuge', because, unlike within standard game theory in which the motives of the players are taken for granted (maximising one's material payoff), 'the whole point [of the current deliberations between Greece's European partners and the new government] is *to forge new motives*' (our emphasis). The main motive of the Greek government

is to implement its social policy agenda, to 'do what is right not as a strategy but simply because it is ... right'. Varoufakis emphasises this commitment not to cross this 'red line', by stating that '[Greece is] *determined* to clash with mighty vested interests in order to reboot Greece and gain our partners' trust. [Greece is] also *determined* not to be treated as a debt colony that should suffer what it must' (our emphasis). He concludes by claiming that:

> 'One may think that this retreat from game theory is motivated by some radical-left agenda. Not so. The major influence here is Immanuel Kant, the German philosopher who taught us that the rational and the free escape the empire of expediency by doing what is right.'

Our point here is not to discuss whether the policy defended by the Greek government is the right one or not, but to offer an analysis of Varoufakis position in terms of our model of preferences. For sakes of clarity, we will refer to the Greek government as agent A and to the European Union as B. The negotiation can be roughly described as a trade-off between (i) the financial aid A could get from B on the one hand, and (ii) the implementation of the social policies promised by A on the other hand. The core issue is that B agrees to offer a financial aid to A only if A does not implement its social policy, although nobody wants A to default. In this article, Varoufakis tries to convince B that A is truthful when claiming that what matters for A is not its 'material payoff' (i.e. the payment of A's debt), but another motive (the help A may provide to its people). Indeed, if B believes that A is simply bluffing, in the sense that A pretends to want to implement its social policy under any circumstances (although A would prefer to implement a austerity policy so as to get B's aid), then A's threat of implementing the social policy is not credible. On the other hand, if A *truly* wants to implement the social policy (and even convince itself that this is the *right* thing to do), then A's threat becomes credible, and B is likely to offer its aid.

It must clearly be noticed that, in the current situation, A is probably *not* bluffing: A genuinely decided to follow another objective than its initial objective in terms of material payoff. The fact that A may *in fine* benefit from B's aid without having to break its electoral promise is only a fortunate by-product of its new preferences. A's commitment to respect its policy agenda is therefore not a bluff, it is a *rational* commitment. It is therefore possible that A, by showing its determination to do 'what is right', chose to 'forge new motives' as a means to satisfy its primary objective (A's new preferences are therefore well *strategic* preferences), but it is also possible that, in line with a more Kantian argument, A deliberately chose its new preferences as an

end in itself (cf. our discussion in chapter 4 on contractarianism vs. contractualism).

Throughout the rest of this thesis, we will consider that one's strategic preferences are purely instrumental: *ex ante*, each individual has a well-identified objective (her material payoff), and chooses her optimal commitment so as to satisfy this objective. We will therefore systematically consider that one's strategic preferences are only valuable as a means to satisfy one's material payoff. In particular, we will not investigate whether one's strategic preferences may progressively become one's material payoff. So as to illustrate this point, think for instance of a philanthropist who helps the needy, because it is her moral duty (and not merely by empathy): she is then satisfying preferences that are different from her material payoff. She may then progressively find 'inner satisfaction in spreading joy' (Kant, 1785), implying that her strategic preferences progressively become her material payoff. She is then acting *in accordance with*, rather than *from*, her duty. Considering this kind of preference evolution would allow us to develop a more general model of preference formation, but this is however beyond the scope of the present work[2].

### 5.2.2 Preliminaries

Let $N = \{1, \dots, n\}$ denote the set of players, with $n \geq 2$. $X = \prod_{i \in N} X_i$ denote the set of pure strategy profiles where each set $X_i \subset \mathbb{R}$ denote the strategy space of player $i$. The material payoff of a player $i \in N$ is given by a function $\Pi_i : X \mapsto \mathbb{R}$, $\forall i \in N$. We assume that players may present interdependent preferences, i.e. that their utility function $U_i : X \mapsto \mathbb{R}$ — whose maximisation determines their choice — is a weighted sum of the material payoff functions $\Pi_j(x)$:

$$U_i(x|S) = \sum_{j \in N} \sigma_{ij} \Pi_j(x), \tag{5.1}$$

with $S = \{\sigma_{ij}\}_{i,j \in N} \in \mathbb{R}^{n \times n}$ a set of real parameters. $\sigma_{ij}$ therefore represents the weight player $i$ gives to player $j$ in her utility function, and its sign indicates whether player $i$ tries to cooperate or not with player $j$. $\Pi_i$ therefore measures the

[2]The model of preference formation could be the following: in period 1, for an initial material payoff $\Pi^1$, my strategic preferences are $U^1$. $U^1$ then progressively becomes my new objective in period 2, i.e. we have $\Pi^2 = U^1$. I am then likely to choose new strategic preferences $U^2$ so as to satisfy my new material payoff $\Pi^2$, which will in turn determine my material payoff in the next periods. The main interest of this kind of approach is that it would not require a notion of metapreferences: the criterion for selecting one's preferences is indeed defined as the strategic preferences of the previous period.

ultimate objective of player $i$ (her material payoff), while $U_i$ represents her strategic preferences, i.e. the optimal interdependent preferences player $i$ should choose so as to maximise *in fine* her material payoff $\Pi_i$.
For any game in normal form $\Gamma = \langle N, X, \Pi \rangle$, we define a two-stage game $\Gamma^*$ as follows:

- in the second stage game $\Gamma_2(S) = \langle N, X, U(.|S) \rangle$, player $i \in N$ chooses a strategy $x_i \in X_i$ so as to maximise her utility function $U_i(x|S)$,
- in the first stage game $\Gamma_1 = \langle N, \mathbb{R}^{n \times n}, V \rangle$, player $i \in N$ chooses a vector of real parameters $S_i = \{\sigma_{i1}; \ldots; \sigma_{in}\}$ so as to maximise her indirect payoff function $V(S) = \Pi(\bar{x}(S))$, with $\bar{x}(S)$ a Nash equilibrium of $\Gamma_2(S)$.

We assume that $\Pi_i$ is a $C^3$ function $\forall i \in N$. Furthermore, we assume that, $\forall S \in \mathbb{R}^{n \times n}$, $\Gamma(S)$ has a unique Nash equilibrium in pure strategies $\bar{x}(S)$, i.e. $\exists! \bar{x}(S) \in X$ such that, $\forall i \in N$:

$$\frac{\partial U_i}{\partial x_i}(\bar{x}(S)|S) = 0, \tag{5.2}$$

$$\frac{\partial^2 U_i}{\partial x_i^2}(\bar{x}(S)|S) < 0. \tag{5.3}$$

This last assumption allows us to considerably alleviate the presentation of our results, although it is not necessary. We could alternatively assume that players are able to use mixed strategies: since the game $\Gamma_2(S)$ is continuous by construction, we know that there necessarily exists at least a Nash equilibrium in mixed strategies. Furthermore, in the case of multiple Nash equilibria, we could define a similar notion of equilibrium than the one we develop in the next section, but by considering multiple subgames for $\Gamma_2$. The two main results of this chapter would also remain unchanged: the demonstration of proposition 1, according to which the players generally have an incentive in presenting strategic preferences different from their material payoff, could indeed easily be extended to a more general framework, while proposition 3, according to which the players choose cooperative (resp. aggressive) preferences in symmetric supermodular (submodular) games is proven under conditions that would ensure the existence of a unique Nash equilibrium in the second stage game (we indeed assume a strong form of diagonal dominance of the Jacobian matrix of marginal utilities).

We introduce the following notations:

- The partial derivatives of $\Pi_i : X \mapsto \mathbb{R}$ are denoted:

$$\Pi_i^{jk}(x) = \frac{\partial^2 \Pi_i}{\partial x_j \partial x_k}(x_1; \ldots; x_n). \tag{5.4}$$

- $J(S) \in \mathbb{R}^{n \times n}$ denote the Jacobian matrix of the marginal utilities evaluated at the Nash equilibrium of $\Gamma_2(S)$:

$$J(S) = \{U_i^{ij}(\bar{x}(S))\}_{i,j \in N} \tag{5.5}$$

- For a $n \times n$ matrix $S \in \mathbb{R}^{n \times n}$, $S_{ij}$ denotes a $(n-1) \times (n-1)$ matrix that results from deleting row $i$ and column $j$ of $S$.

- For a $n \times n$ matrix $S \in \mathbb{R}^{n \times n}$, $C_{ij}^S = (-1)^{i+j} |S_{ij}|$ denotes the $(i; j)$ cofactor of $S$.

The notation for the derivatives also holds for the utility function $U_i$. The game $\Gamma(S)$ is supermodular (respectively submodular) if and only if, $\forall i \in N$:

$$U_i^{ij}(x) \geq (\leq) 0 \qquad \forall x \in X, \quad \forall j \neq i. \tag{5.6}$$

We make the additional assumption that $J(S)$ and its minors $J_{ii}(S)$ are generically non singular $\forall S \in \mathbb{R}^{n \times n}$.

### 5.2.3 Subgame perfect equilibrium of commitment

Suppose that the players are able to choose their own weights $\sigma_{ij}$: they can therefore make strategic commitments, since their choice is determined by the maximisation of their utility function $U_i(x|S)$, and their payoff is determined by their material payoff $\Pi_i(x)$. Since we assumed the existence of a unique Nash equilibrium for $\Gamma_2(S)$, $\forall S \in \mathbb{R}^{n \times n}$, we can define the indirect material payoff $V_i : \mathbb{R}^{n \times n} \mapsto \mathbb{R}$ as follows[3]

$$V_i(S) = \Pi_i(\bar{x}(S)). \tag{5.7}$$

We can now define a notion of equilibrium that embodies the idea of a strategic choice of one's commitments:

[3] We relax the assumption of a unique second stage equilibrium in chapter 7: the approach we will use can be directly transposed to the present situation.

**Definition 1.** *Let* $\Gamma = \langle N, X, \Pi \rangle$ *denote a game in normal form. A strategy profile* $(\bar{x}; \bar{S}) \in X \times \mathbb{R}^{n \times n}$ *is a subgame perfect equilibrium of commitment (SPEC) of* $\Gamma$ *if and only if:*

- $\bar{x} \in X$ *is a Nash equilibrium of* $\Gamma_2(\bar{S})$,
- $\bar{S} \in \mathbb{R}^{n \times n}$ *is a Nash equilibrium of* $\Gamma_1$.

A SPEC is therefore a specific utility function (defined by the degree of interdependence with the other players) and a strategy profile of the initial game such that no player obtains a strictly higher material payoff by changing her strategic commitment $S_i$, i.e. there exists no game $\Gamma_2(S_i; \bar{S}_{-i})$ with $S_i \neq \bar{S}_i$ such that $i$ is better off at the Nash equilibrium of $\Gamma_2(S_i; \bar{S}_{-i})$ than at the Nash equilibrium of $\Gamma_2(\bar{S})$.

### 5.2.4 Illustration

Consider a game $\Gamma = \langle \{1, 2\}, \{\mathbb{R}^+\}^2, \Pi \rangle$, with:

$$\Pi_i(x_1, x_2) = ay + \frac{b}{2}y^2 - \frac{c}{2}x_i^2, \qquad a, c > 0,\ 4b < c, \tag{5.8}$$

with $y = (x_1 + x_2)$ if $b \geq 0$ and $y = \min\{(x_1 + x_2); |a/b|\}$ if $b < 0$ (this last condition ensures that the function $ay + \frac{b}{2}y^2$ is always increasing). $\Gamma$ is a public good game, in which each player chooses a level $x_i$ that generates a collective benefit and an individual cost. We can associate to $\Gamma$ a two-stage game $\Gamma^*$. In the second stage game $\Gamma_2(S)$, the players maximise their utility functions $U_i$:

$$U_i(x_1, x_2|S) = (\sigma_{i1} + \sigma_{i2}) \left[ ay + \frac{b}{2}y^2 \right] - \frac{c}{2} \left[ \sigma_{i1}x_1^2 + \sigma_{i2}x_2^2 \right]. \tag{5.9}$$

We can easily check that $\sigma_{ii} = 0$ cannot be a first stage equilibrium (if $b > 0$, player $i$ chooses $x_i \to +\infty$ and gets her worst possible payoff; if $b < 0$, player $i$ chooses $x_i = |a/b|$ and supports all the costs). We therefore normalise $\sigma_{ii}$ to 1, $\forall i \in N$. The unique Nash equilibrium of $\Gamma_2(S)$ is then:

$$\bar{x}_i(S) = \frac{a(1 + \sigma_{ij})}{c - (2 + \sigma_{12} + \sigma_{21})b}, \qquad \forall i \in N. \tag{5.10}$$

(5.10) gives the optimal effort of each player given the weights they attribute to the other player within their utility function (we can verify that $(\bar{x}_1 + \bar{x}_2) < |a/b|$ when $b < 0$). We now assume that both players are able to directly choose those weights in a precommitment game $\Gamma_1 = \langle N, \mathbb{R}^2, V \rangle$, with $V_i = \Pi_i(\bar{x}(S))$ the indirect payoff function of player $i$:

$$V_i(S) = \frac{a^2(2 + \sigma_{12} + \sigma_{21})((1 - 2\sigma_{12} - \sigma_{12}^2)c - (2 + \sigma_{12} + \sigma_{21})b)}{2(c - (2 + \sigma_{12} + \sigma_{21})b)}. \tag{5.11}$$

We can then compute the Nash equilibrium of the first stage game (we look here for an interior solution for $\bar{x}$):

$$\bar{\sigma}_{ij} = \frac{c - 2b - \sqrt{c(c - 4b)}}{2b}, \qquad \forall i \in N. \tag{5.12}$$

We therefore obtain a unique SPEC $(\bar{x}; \bar{S}) \in \{\mathbb{R}^+\}^2 \times \mathbb{R}^2$:

$$\begin{cases} \bar{x}_i = \dfrac{2ab - c + \sqrt{c(c - 4b)}}{2b\sqrt{c(c - 4b)}}, & \forall i \in N, \\ \bar{\sigma}_{ij} = \dfrac{c - 2b - \sqrt{c(c - 4b)}}{2b}, & \forall i \in N,\ j \neq i. \end{cases} \tag{5.13}$$

We can notice that the weights $\bar{\sigma}_{ij}$ at equilibrium have the same sign than the parameter $b$, i.e. that players will play cooperatively — and produce a higher output than at Nash equilibrium — if and only if the game is supermodular, even if the cooperation is not full (we have indeed $\bar{\sigma}_{ij} < 1$). We can finally check that the profits of both players are superior to the profits at Nash equilibrium if and only if the game is supermodular. Conversely, for a game with a concave benefit function ($b < 0$), the players will be more competitive at the first stage equilibrium and will therefore get a lower outcome. Indeed, with a concave benefit function (i.e. with strategic substitutes), each player has an incentive to 'blackmail' the other one — i.e. to unilaterally decrease her own output — in order to force the other player to increase her output. Since both players have the same reasoning, they enter in a vicious circle and end up with a deteriorated situation.

This illustration highlights the possible connection between supermodularity and the endogenous formation of cooperative preferences. We can find a similar result within the literature on the indirect evolutionary approach: Bester and Güth (1998) for instance argue on the one hand that altruism is evolutionary stable in some games

presenting strategic complementarities, while Bolle (2000) and Possajennikov (2000) notice on the other hand that relaxing this assumption will lead to the evolutionary stability of spite and anti-social motives. We now generalise this result and determine the general form of the parameters $\sigma_{ij}$ at the equilibrium of the first stage game.

## 5.3 Optimal interdependent preferences

We firstly introduce the notions of Stackelberg best reply and payoff functions, and then determine the optimal weights $\bar{\sigma}_{ij}$.

### 5.3.1 Stackelberg best reply and payoff functions

Before presenting our main result, we need to introduce the notion of *Stackelberg best reply function* and *Stackelberg payoff function*. The Stackelberg best reply function of player $j$ is her best reply to the strategy of player $i$, knowing that the players $k \neq i, j$ are maximising their utility: it is the reply function a Stackelberg leader will use so as to predict the behaviour of her opponents. The Stackelberg payoff function is simply the material payoff of player $i$ that integrates the Stackelberg best reply functions of the other players, and whose maximisation determines the strategy chosen by a Stackelberg leader with $(n-1)$ followers.

**Definition 2.** *The function $f_j : S \times X_i \mapsto X_j$ is the Stackelberg best reply function of player $j$ for $S \in \mathbb{R}^{n \times n}$ if and only if, $\forall x_i \in X_i$:*

$$U_j(f_{-j}(x_i|S); f_j(x_i \mid S)|S) \geq U_j(f_{-j}(x_i|S); x_j \mid S), \qquad \forall x_j \in X_j, \tag{5.14}$$

*with $f_k : S \times X_i \mapsto X_k$ the Stackelberg best reply function of player $k$ for $S \in \mathbb{R}^{n \times n}$, $\forall k \neq i, j$.*

The incomplete strategy profile $x_{-i} = \{f_1(x_i|S); \ldots; f_n(x_i|S)\} \in X_{-i}$ can be interpreted as the Nash equilibrium of the game $\Gamma_2(S) \setminus x_i = \langle N \setminus i, X_{-i}, U_{-i} \rangle$, which is function of a 'parameter' $x_i$. The existence of a Nash equilibrium in $\Gamma_2(S)$ implies that the best reply functions $f_j(.|S)$ are defined on a non empty subset of

$X_i$. Indeed, if it was not the case, then a second stage equilibrium could not exist, since $\bar{x} \in X$ is a Nash equilibrium of $\Gamma_2(S)$ if and only if:

$$f_j(\bar{x}_i|S) = \bar{x}_j, \qquad \forall i, j \in N. \tag{5.15}$$

For the same reasons motivating our assumption that each second stage game has a unique Nash equilibrium, we assume that there always exists a unique function $f_j : X_i \mapsto X_j$, $\forall i, j \in N$. The reasoning supporting proposition 1 can indeed easily be extended to a more general framework with several functions $f_j$ (their existence being ensured by the existence of a Nash equilibrium in mixed strategies for each game $\Gamma_2(S)$), and the conditions allowing us to establish the relation between supermodularity and the formation of cooperative preferences would typically imply the uniqueness of the Stackelberg best reply function.
We can now define the Stackelberg function:

**Definition 3.** *Let $f_j(x_i)$ denote the Stackelberg best reply function of player $j$ for $S$. The function $\Psi_i : X_i \mapsto \mathbb{R}$ is the Stackelberg function of player $i$ if and only if:*

$$\Psi_i(x_i|S) = \Pi_i(f_1(x_i|S); \ldots; f_n(x_i|S)). \tag{5.16}$$

As a preparation for our propositions we show the following lemmas (the proofs are provided in appendix):

**Lemma 1.** *Let $f_j(x_i|S)$ be the Stackelberg best reply function of player $j$ for $S$. We have:*

$$\frac{\partial f_j}{\partial x_i}(x_i|S) = \frac{C_{ij}^{J(S)}}{C_{ii}^{J(S)}}, \tag{5.17}$$

*with $C_{ij}^{J(S)}$ the $(i;j)$ cofactor of $J(S)$, the Jacobian matrix of the marginal utility functions $U_i^i(x|S)$, evaluated at the Nash equilibrium of $\Gamma_2(S)$.*

**Lemma 2.** *If $\forall j, k \neq i$:*

*(i)* $\left|U_i^{ii}(\bar{x}(S)|S)\right| > (n-1)\left|U_i^{ij}(\bar{x}(S)|S)\right|$,

*(ii)* $\left|U_i^{ik}(\bar{x}(S)|S)\right| < (n-1)\left|U_i^{ij}(\bar{x}(S)|S)\right|$,

*then:*

$$sign\left(\frac{\partial f_j}{\partial x_i}(x_i|S)\right) = sign\left(U_j^{ji}(\bar{x}(S)|S)\right), \qquad \forall j \neq i. \tag{5.18}$$

**Lemma 3.** *If* $\forall j, k \neq i$*:*

*(i)* $\left|U_i^{ii}(\bar{x}(S)|S)\right| > (n-1)\left|U_i^{ij}(\bar{x}(S)|S)\right|$,

*(ii)* $\left|U_i^{ik}(\bar{x}(S)|S)\right| < (n-1)\left|U_i^{ij}(\bar{x}(S)|S)\right|$,

*then:*

$$\sum_{j\neq i}\left|C_{ij}^{J(S)}\right| < \left|C_{ii}^{J(S)}\right|, \tag{5.19}$$

$$\iff \sum_{j\neq i}\left|\frac{\partial f_j}{\partial x_i}\right| < 1. \tag{5.20}$$

Lemma 1 provides the expression of the first order derivative of the Stackelberg best reply function $f_j(x_i)$ as a function of the second order derivatives $U_i^{ij}$. We can then show that the sign of $\frac{\partial f_j}{\partial x_i}$ is the same than $U_j^{ji}(\bar{x}(S)|S)$, when conditions (i) and (ii) are verified (lemma 2). Condition (i) means that $j$'s impact on $i$'s marginal utility is relatively low compared to $i$'s impact on her own marginal utility (this is a strong form of row diagonal dominance), and (ii) the cross derivatives $U_i^{ij}$ and $U_i^{ik}$ are relatively 'close' in absolute value $\forall j, k \neq i$, i.e. there is no player $j$ with a significantly higher importance from $i$'s perspective. Those conditions are typically verified for public good games and two-player games with $\left|U_i^{ii}\right| > \left|U_i^{ij}\right|$. Lemma 3 states that, under the same conditions, the sum of the $(i;j)$ cofactors, $\forall j \neq i$, is lower than the principal minor of $J(S)$.

We now determine the expression of the weights $\bar{\sigma}_{ij}$ at the Nash equilibrium of $\Gamma_1$, and determine their sign: we will then be able to define a class of games in which players endogenously adopt cooperative preferences, or conversely try to maximise the difference between their payoff and the payoff of their opponents.

### 5.3.2 Optimal weights

Let $\Gamma = \langle N, X, \Pi \rangle$ be a game in normal form, and $\Gamma^*$ its associated two-stage game. We firstly determine the conditions under which it is rational for all the players to choose null weights $\bar{\sigma}_{ij}$, $i \neq j$, i.e. all the players prefer to maximise their material payoff rather than adopting interdependent preferences:

**Proposition 1.** *Let $\bar{x} \in X$ be the Nash equilibrium of $\Gamma$, and $I_n$ a matrix in $\mathbb{R}^{n \times n}$ such that $\sigma_{ij} \neq 0$ iff $i = j$. $(\bar{x}; I_n)$ is a SPEC of $\Gamma$ if and only if:*

*(i) either $\Pi_i^j(\bar{x}) = 0$, $\forall i, j \in N$,*

*(ii) or $\Psi_i^i(\bar{x}) = 0$, $\forall i$ such that $\exists j \neq i$ for which $\Pi_i^j(\bar{x}) \neq 0$.*

Proposition 1 states that, unless the interests of the players are perfectly aligned or opposed (in the sense that maximising one's payoff implies also maximising or minimising the payoffs of all the other players — as implied by condition (i)), or that no-one can benefit from a first mover advantage (condition (ii)), there is at least one player who will be better off by choosing a non null weight $\sigma_{ij}$. The intuition behind this result is the following: in a strategic interaction with payoff maximisers, the highest payoff I can achieve is my Stackelberg payoff, i.e. the payoff I would get if I was able to play before the others (suppose here for the sake of argument that a player with a first mover advantage can always obtain the Nash payoff — this means for instance that, in a zero-sum game, a Stackelberg leader could play in mixed strategies). If I have the opportunity to choose strategic preferences different from my material payoff, then I can manipulate the Nash equilibrium of the game such that the strategy that satisfies my strategic preferences actually satisfies my Stackelberg payoff, i.e. such that the Nash equilibrium with strategic preferences corresponds to the Stackelberg equilibrium with my initial material payoff.
We can now provide the expression of the optimal weights:

**Proposition 2.** *$(\bar{x}; \bar{S})$ is a SPEC of $\Gamma$ if, $\forall i, j \in N$:*

$$\bar{\sigma}_{ij} = \frac{\Pi_i^j}{\Pi_j^i}(\bar{x}) \frac{\partial f_j}{\partial x_i}(\bar{x}_i | \bar{S}). \tag{5.21}$$

Proposition 2 gives the expression of the optimal weights a player $i$ should give to the other players so as to maximise her material payoff (we can check that $i$ necessarily maximises her own payoff, since $\bar{\sigma}_{ii} = 1$). This condition is not necessary, since — as shown in the proof — the vector $\bar{S}_i$ is determined by a single equation. Although other specifications were possible, we chose here to define $\bar{\sigma}_{ij}$ as a function of the Stackelberg best reply of player $j$ when $i$ is the leader, since it captures the idea that the attitude of $i$ towards $j$ fundamentally depends on the way $j$ reacts when $i$ changes her strategy.

The weights $\bar{\sigma}_{ij}$ are therefore chosen such that satisfying my strategic preferences is formally equivalent to maximising my Stackelberg function. It can then be interesting to determine under which conditions the choice of one's preferences implies a more cooperative or competitive behaviour with the other players, i.e. to determine the sign of the optimal weights $\bar{\sigma}_{ij}$. Thanks to lemmas 1 and 2, we can see that the sign of $\bar{\sigma}_{ij}$ is determined by the sign of $U_j^{ji}$ (they have the same sign if and only if $sign(\Pi_i^j(\bar{x}(\bar{S}))) = sign(\Pi_j^i(\bar{x}(\bar{S})))$, which is for instance the case in public good games or Cournot oligopoly). It means that player $i$ will cooperate with player $j$ if and only if there is a strategic complementarity between $i$ and $j$ in the game $\Gamma_2(\bar{S})$. This implies in particular that, in supermodular games, players have an interest in presenting cooperative preferences, since this will generate a positive best reply from the other players: cooperating is therefore beneficial because it gives an incentive to other players to reciprocate. On the contrary, games with strategic substitutes will generate more competitive behaviours, the players having an incentive in maximising the difference between the payoffs rather than their sum.
Note however that proposition 2, lemma 1 and lemma 2 are not sufficient to ensure that players will necessarily cooperate if the initial game $\Gamma$ is supermodular: the condition holds only for the resulting game $\Gamma_2(\bar{S})$. It is in fact not impossible that there exists a SPEC in a supermodular game such that all players present negative $\sigma_{ij}$ (it can be consistent if the resulting game is submodular): there can therefore exist Nash equilibria in the first stage game that create an artificial competition between the players, although the initial game was supermodular. We can however notice that the only reason for $i$ to compete with $j$ is that $j$ competes at equilibrium with $i$: it seems quite unlikely that players will effectively converge to such an equilibrium.
A corollary of those results is that, if only one player $i$ is able to make strategic commitments (as in a game with a Stackelberg leader), then the strategy chosen by this Stackelberg leader would correspond to the satisfaction of cooperative (resp. competitive) preferences in supermodular (submodular) games: Stackelberg leader-

ship therefore leads to a greater cooperation in supermodular games, and a greater competition in submodular games.
We now show that the connection between the supermodularity of the initial game $\Gamma$ and positive $\bar{\sigma}_{ij}$ holds for symmetric games.

### 5.3.3 Symmetric games

We focus here on symmetric games to establish a direct connection between supermodularity and the choice of cooperative preferences at the equilibrium of $\Gamma_1$.

**Definition 4.** *A game in normal form* $\Gamma = \langle N, X, \Pi \rangle$ *is symmetric if and only if:*

- $X_i = X_j$, $\forall i, j \in N$,
- *for any permutation* $s : N \mapsto N$*:*

$$\Pi_i(x_1; \ldots; x_i; \ldots; x_n) = \Pi_{s(i)}(x_{s(1)}; \ldots; x_{s(i)}; \ldots; x_{s(n)}). \tag{5.22}$$

We have the following proposition (proof in appendix):

**Proposition 3.** *Let* $\Gamma$ *be a symmetric game. If* $\forall j, k \neq i$*:*

*(i)* $\left|U_i^{ii}\right| > (n-1)\left|U_i^{ij}\right|$,

*(ii)* $\left|U_i^{ik}\right| < (n-1)\left|U_i^{ij}\right|$,

*(iii)* $\left|\Pi_j^{ji}\right| \geq \left|\Pi_k^{ji}\right|$,

*then a symmetric SPEC* $(\bar{x}; \bar{S})$ *verifies:*

$$sign\left(\bar{\sigma}_{ij}\right) = sign\left(\Pi_j^{ji}(\bar{x}(S))\right), \qquad \forall j \neq i. \tag{5.23}$$

Proposition 3 states that the connection between supermodularity of the initial game and cooperation in second stage game is true for symmetric $n$-player games, under the assumptions (i) and (ii) introduced in the previous section, and under

the additional assumption that the second order derivative $\Pi_k^{ji}$ (for different $i$, $j$ and $k$) is relatively low in absolute value. This result means that, in a symmetric game, player $i$ will choose to put a positive weight on the material payoff of player $j$ if and only if $\Pi_j^{ji}$ is positive in the initial game: supermodular games will then endogenously generate cooperative preferences. The cooperation is not full, since lemma 3 implies that:

$$\sum_{j \neq i} |\bar{\sigma}_{ij}| < 1, \qquad \forall i \in N. \tag{5.24}$$

This means that player $i$ will never assign a higher weight to the set of the other players compared to her own material payoff in her preferences. On the contrary, games with strategic substitutes will exacerbate the competition between the players and lead to more aggressive behaviours.

## 5.4 Application: climate change negotiations

We now consider the possible implications of our results for the design of public policies.

### 5.4.1 How public policies shape individual preferences

A core argument of the microfoundations program in macroeconomics relies on the Lucas critique, according to which econometric models should integrate individual optimisation behaviours when assessing the implications of economic policies:

> 'This essay has been devoted to an exposition and elaboration of a single syllogism: given that the structure of all econometric model consists of optimal decision rules of economic agents, and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models.' (Lucas, 1976, p.41)

The Lucas critique therefore stresses that public policies may impact individual behaviours, since altering the strategic environment is likely to induce a new optimal behaviour regarding the satisfaction of one's preferences. We can however notice that this interpretation implicitly assumes that the underlying preferences remain identical across policy regimes: individual behaviour may evolve *only* because

the new institutional setting implies a new optimal decision, and not because those underlying preferences may also change. We can for instance think of crowding-out effects, and the impact of the introduction of pecuniary incentives on pro-social motives. Several works highlighted that those kinds of policies can have counter-productive effects, since the intrinsic motivation for a social objective disappears and is replaced by an extrinsic motivation: individual preferences are then likely to change, and the satisfaction of those new preferences can lead to a worse situation than before the implementation of the policy (see for instance Titmuss (1970) on blood donation, Frey and Oberholzer-Gee (1997) on the willingness to accept a NIMBY project, and also Ostman (1998) and Cardenas et al. (2000) on the management of common-pool resources).

Within our framework, we can state — paraphrasing Lucas — that given that individual preferences consist of optimal decision rules in the first stage game, and that optimal decision rules vary systematically with changes in the structure of the strategic environment, then any change in policy will systematically alter individual preferences. The design of public policies should therefore integrate the possibility that the players will adapt their preferences. Our results imply that, in games characterised by strategic substituability, such as public good games with a concave benefit function, the players have an incentive to become more aggressive: it is therefore possible that the total contribution progressively decreases with the emergence of more competitive preferences, leading *in fine* to a deteriorated situation (worse than the initial Nash equilibrium). An interesting policy recommendation in this kind of situation would be to change the incentives of the initial game such that the players are not tempted any more to adopt such preferences. A solution to promote cooperation would then be to transform the game into one presenting strategic complementarities: this should indeed endogenously lead to more cooperative behaviours.

The aim of this section is to illustrate this point by studying climate change negotiations: we can see that the different solutions suggested up to now consist in designing economic incentives to reduce greenhouse gas emissions (with for instance the Kyoto Protocol or the European Union Emissions Trading Scheme). Those approaches however keep the strategic substituability of the initial game of pollution abatement, since there is a perfect substituability between the emission of two countries $i$ and $j$, and no coordination mechanism between the choice of the different countries. This may in turn give an incentive to the countries to adopt more aggressive positions in international negotiations: they can indeed threaten the other countries to lower their contribution, so as to force them to provide a greater effort. We can indeed reasonably assume that the countries are able to make strategic

commitments, since international negotiations are not only a matter of economic interest, but also of political influence. Efficient international agreements should therefore build a system of incentives that give a coordination structure to the game of pollution abatement, by relying for instance on the adoption of technological standards and trade sanctions to punish the countries not respecting the agreement (this argument is in line with the recommendations of Barrett (2003, 2007) concerning the design of international agreements).

### 5.4.2 Model

We consider two identical countries $i \in N$, in which a firm produces and sells a consumption good in quantity $q_i$. There is no international trade, and the national firm takes the national price $p_i$ as given. The production of the final good generates a pollution $D(q_1 + q_2)$ that negatively affects both countries. Each country can tax the production of its firm (tax rate of $\tau_i$ per unit of production), and is therefore able to indirectly set its level of production. The countries are facing a public good game: they indeed choose their level of production (*via* their national regulation) so as to maximise their national payoff $\Pi_i$, knowing that this generates a national gain (in terms of surplus) but a global loss (pollution). This leads in turn to an over-production and a Pareto-dominated Nash equilibrium. We now suppose that the countries want to implement an international agreement in order to maximise the global payoff $\Pi = \Pi_1 + \Pi_2$, knowing that — once the agreement is signed — both countries will choose their level of production so as to maximise $\Pi_i$.
In this model, firms are in perfect competition, the two countries play a game, and they try to reach an agreement from the social planner's perspective. The objective of this illustration is to compare several alternatives of international systems and to argue in favour of systems creating a game of coordination between the countries rather than keeping the strategic substituability of the initial game. For convenience, we assume that the countries can implement an international tax system such that no fraud is possible, and the funds are collected by an international fund (we can assume for instance that those funds are then used to indemnify the victims of the pollution). This model offers a very simple picture of the current negotiations on climate change: the national productions generate the emission of greenhouse gases, and the different countries try to establish an international system so as to reduce the environmental damage of their production. The international fund we model can be assimilated to the Green Climate Fund, which is funded by the developed countries emitting a higher quantity of greenhouse gases. We suggest now comparing two main scenarios:

- International carbon tax (ICT): each country must pay a constant tax $t_{ICT}$ per unit of pollution emitted
- Trade sanction (TS): the country with the less demanding regulation must pay a tax $t_{TS}(\tau_1;\tau_2)$ to the other country per unit of pollution emitted; this tax depends on the difference between national taxes

ICT seems to correspond to the ideal solution from an economic perspective: it enables to internalise the negative externality of the production, and then to reach a Pareto optimal outcome (by determining the adequate level of taxation). In the second scenario, although there is no international trade, the situation can be related to a mechanism of trade sanction: if a country is in a situation of environmental dumping, then its partners may impose additional taxes on the goods exported by this country (such that no firm can be eventually advantaged by a less restricting regulation). The situation here is relatively similar, since the country directly pays to the other an additional tax in case of environmental dumping. Formally, the material payoff of country $i$ is the following:

$$\Pi_i(q|ICT) = CS_i(q_i) + \pi_i(q_i) + \tau_i q_i - D(q_1+q_2) - t_{ICT} q_i, \tag{5.25}$$
$$\Pi_i(q|TS) = CS_i(q_i) + \pi_i(q_i) + \tau_i q_i - D(q_1+q_2) - t_{TS}(\tau_j - \tau_i) q_i, \tag{5.26}$$

with $CS_i$ the consumer surplus and $\pi_i$ the profit of the firm from country $i$. We assume a linear inverse demand, convex costs for the firms, and a convex damage function:

$$p_i = a - bq_i, \tag{5.27}$$
$$\pi_i = (p_i - \tau_i)q_i - \frac{c}{2}q_i^2, \tag{5.28}$$
$$D(q) = \frac{\delta}{2}(q_1+q_2)^2. \tag{5.29}$$

Since the firms are in perfect competition on each national market, we can easily compute the consumer surplus $CS_i$, as well as the production $q_i$ as a function of $\tau_i$:

$$CS_i(q_i) = \frac{b}{2}q_i^2, \tag{5.30}$$
$$CS_i(q_i) + \pi_i(q_i) + \tau_i q_i = aq_i - \frac{b+c}{2}q_i^2, \tag{5.31}$$
$$\text{with} \qquad q_i = \frac{a-\tau_i}{b+c} \tag{5.32}$$

Without international agreement, the material payoff of each country is therefore:

$$\Pi_i(q) = aq_i - \frac{b+c}{2}q_i^2 - \frac{\delta}{2}(q_1+q_2)^2, \tag{5.33}$$

and the Nash equilibrium, $\forall i \in N$:

$$\begin{cases} \bar{q}_i = \dfrac{a}{b+c+2\delta}, \\ \bar{\tau}_i = \dfrac{2a\delta}{b+c+2\delta}. \end{cases} \tag{5.34}$$

Both countries are therefore producing too much, since the social optimum (maximising the sum of material payoff) is, $\forall i \in N$:

$$\begin{cases} \tilde{q}_i = \dfrac{a}{b+c+4\delta}, \\ \tilde{\tau}_i = \dfrac{4a\delta}{b+c+4\delta}. \end{cases} \tag{5.35}$$

Furthermore, the game is here submodular: players have therefore an incentive to adopt competitive preferences and negative $\sigma_{ij}$. The unique SPEC of the game is:

$$\begin{cases} \bar{\sigma}_{ij} = \dfrac{\sqrt{(b+c)(b+c+4\delta)} - b - c - 2\delta}{2\delta}, \\ \bar{q}_i = \dfrac{a}{\sqrt{(b+c)(b+c+4\delta)}}. \end{cases} \tag{5.36}$$

Both countries try to get the upper hand on the other, and end up *in fine* with a deteriorated situation, in which they both produce more than at Nash equilibrium.

### 5.4.3 Scenario ICT

Consider firstly that both countries implement ICT: they therefore pay to a third party a tax $t_{ICT}$ per unit of pollution. Their material payoff is now:

$$\Pi_i(q) = aq_i - \frac{b+c}{2}q_i^2 - \frac{\delta}{2}(q_1+q_2)^2 - t_{ICT}q_i. \tag{5.37}$$

Suppose firstly that the countries agree on a 'naive' tax, i.e. a tax that does not take into account the possibility that players may make strategic commitments,

once the agreement is signed. In the absence of strategic commitment, the level of the tax that allows the countries to reach the social optimum (5.35) is:

$$t_{ICT,n} = \frac{2a\delta}{b+c+4\delta}. \tag{5.38}$$

If the countries agree to implement this international tax, then they have the adequate incentives to reach the optimal production $\tilde{q}$, given their initial preferences. We can now notice that this policy keeps the submodularity of the initial game: a player can therefore benefit from a strategic commitment such that she eventually produces $\hat{q} > \tilde{q}$, since the best reply of the other country will be to increase her effort (by reducing her production). In particular, since the implementation of a tax per unit of production does not affect the cross derivatives $\Pi_i^{ij}$, the players will choose the exact same weights as previously, and the tax (5.38) will not give the adequate incentives to reach the social optimum.
Suppose therefore that the countries anticipate that they will make strategic commitments once the agreement is signed. The unique SPEC is then:

$$\begin{cases} \bar{\sigma}_{ij} = \dfrac{\sqrt{(b+c)(b+c+4\delta)} - b - c - 2\delta}{2\delta}, \\ \bar{q}_i = \dfrac{(a - t_{ICT})(b+c+\delta)}{(b+c)^2 + 2\delta(b+c)}. \end{cases} \tag{5.39}$$

The optimal tax rate is therefore:

$$t^*_{ICT} = \frac{2a\delta(1-\bar{\sigma})}{b+c+4\delta}, \tag{5.40}$$

$$\Leftrightarrow \quad t^*_{ICT} = a\left[1 - \sqrt{\frac{b+c}{b+c+4\delta}}\right], \tag{5.41}$$

which is strictly higher than the naive tax $t_{ICT,n}$.

An international agreement must therefore take into account the possibility *ex post* for the countries to benefit from strategic commitments. The question that arises then is to know whether the countries are likely to make the optimal strategic commitment: a crucial condition for choosing an optimal commitment is indeed that we anticipate that the others know that we will respect our commitment. If we do not believe that the other is sufficiently rational to play the first stage game[4], or

[4] We will discuss the epistemic conditions required for players to directly choose their preferences in the first stage game in chapter 6.

alternatively that preferences are not directly chosen, but are the result of evolutionary pressures, then it is not certain that $t^*_{ICT}$ will be well-suited. In addition of preventing the players from adopting competitive strategic commitments, we should also implement a tax system such that the optimal policy does not depend on the level of $\sigma_{ij}$, on which the social planner has no direct control.
Although an international tax may lead *in fine* to the social optimum, we can notice that the players necessarily adopt aggressive preferences at equilibrium — which may be an undesirable property (in a non-welfarist perspective) of international relations. Furthermore, such a system is highly sensitive to the propensity of the players to respect their optimal commitment: if one player does not keep her optimal commitment, either because she is not rational enough, or because she does not think the other will keep her commitment, or because preferences evolve over time *via* an evolutionary dynamics, then the tax will probably not be well adapted.

### 5.4.4 Scenario TS

Consider now the scenario TS. Assume that $\tau_i < \tau_j$: the country $i$ (with the lowest national tax) must pay a tax $t_{TS}(\tau_j - \tau_i)$ per unit of pollution, and country $j$ collects this tax within the limits of its own production. The residual is collected by the international organisation[5]. Within this scenario, the material payoff of country $i$ is, $\forall i \in N$:

$$\Pi_i(q) = aq_i - \frac{b+c}{2}q_i^2 - \frac{\delta}{2}(q_1+q_2)^2 - t_{TS}(\tau_j - \tau_i)q_i, \tag{5.42}$$

$$\iff \Pi_i(q) = aq_i - \frac{b+c}{2}q_i^2 - \frac{\delta}{2}(q_1+q_2)^2 - t_{TS}(b+c)(q_i - q_j)q_i. \tag{5.43}$$

As previously, consider firstly that the countries agree on a naive tax rate, that does not take into account the possibility for the countries to make strategic commitments *ex post*. The expected Nash equilibrium if scenario TS is chosen is then:

$$q_i = \frac{a}{(b+c)(1+t_{TS}) + 2\delta}, \qquad \forall i \in N, \tag{5.44}$$

$$\tau_i = \frac{a(2\delta + (b+c)t_{TS})}{2\delta + (b+c)(1+t_{TS})}, \qquad \forall i \in N. \tag{5.45}$$

[5]We adopt this specific framework to be consistent with a scenario of trade sanctions, in particular if one country is bigger than the other: the funds collected by $j$ must indeed be in a scale consistent with its own production.

So as to reach the social optimum (5.35), we should therefore implement the following tax rate:

$$t_{TS} = \frac{2\delta}{b+c}. \tag{5.46}$$

We can now notice that, although this tax does not take into account the possibility that countries may strategically choose their own preferences, the game with the payoff function described in (5.43) is supermodular if and only if $t_{TS} > \frac{\delta}{b+c}$, which is verified for the optimal tax defined in (5.46). The game with trade sanctions is therefore supermodular, and should lead to the emergence of cooperative behaviours at the SPEC. The optimal weights of each country are indeed, $\forall i \in N$:

$$\bar{\sigma}_{ij} = \frac{\sqrt{(b+c)^2(t_{TS}+1)^2 + 4\delta(b+c)(t+1)} - (b+c)(t_{TS}+1) - 2\delta}{2\delta}, \tag{5.47}$$

which is positive when $t_{TS} > \frac{\delta}{b+c}$, i.e. when the game in scenario TS is supermodular. Since this condition is verified for the optimal tax rate (5.46), we know that players will tend to be more cooperative with scenario TS.
We saw that the naive tax in the case of scenario ICT was not adapted when players were making a strategic commitment (the optimal level of tax indeed directly depended on the level of $\sigma_{ij}$). A crucial difference between scenarios ICT and TS is that the naive tax implemented with TS still remains optimal at the SPEC. We can indeed show that the level of tax $t_{TS}$ that maximises social welfare does not depend on $\sigma_{ij}$ when $\sigma_{12} = \sigma_{21}$. We have indeed in this situation the optimal production for $i$:

$$q_i = \frac{a}{(b+c)(1+t_{TS}(1-\sigma)) + 2\delta(1+\sigma)}, \tag{5.48}$$

which is equal to the social optimum $\tilde{q}$ if and only if:

$$t_{TS} = \frac{2\delta}{b+c}. \tag{5.49}$$

This property allows us to tackle the two issues faced in scenario ICT, (i) that countries were likely to adopt aggressive preferences, and (ii) that, from a more practical perspective, the level of the tax depended on the likelihood for both countries to implement their optimal commitment. Firstly, adopting a mechanism of trade sanctions is likely to generate cooperative preferences between countries, since their interests are now aligned: if a country increases its level of effort (by

increasing $\tau_i$, and therefore diminishing $q_i$) then the cost is shared among both countries. The country with the lower effort must indeed now pay a compensation to the other country, and has now a new incentive to increase its own effort: in addition to diminishing the environmental damage, the country will also stop paying the other country. Secondly, as long as the preferences of the country evolve in a similar way (and therefore $\sigma_{12} = \sigma_{12}$), the tax rate $t_{TS}$ given by (5.46) remains optimal over time. It is therefore possible to implement a naive tax rate, since this level of taxation will still be optimal *ex post*, once the countries have made symmetric commitments.

International agreements relying on a mechanism of trade sanctions rather than an international tax are more likely to succeed, since they align the interests of the different countries: since the game is supermodular (both countries indeed know that increasing one's effort will increase the incentives for the other to increase its own effort), cooperative behaviours are likely to emerge, unlike with scenario ICT, in which the initial submodularity of the game is conserved, leading to the emergence of aggressive behaviours. We can indeed notice that the weights $\bar{\sigma}_{ij}$ chosen at the SPEC are increasing with $t_{TS}$, and that $\lim_{t_{TS}\to+\infty} \bar{\sigma}_{ij} = 1$: this means that by increasing the level of sanction, the players will naturally converge to cooperative preferences and directly maximise the global welfare.
A last interesting property of scenario TS is that, at a symmetric equilibrium, there is no transfer between the countries or with the international fund (both countries have indeed the same level of production). While the collect of the tax with scenario ICT is likely to generate additional costs, it is possible to achieve the same results in terms of individual incentives without having to make any transfer between countries.

## 5.5 Conclusion

It is generally assumed in game theory that players have a fixed material payoff function and that their ultimate objective is to achieve the highest possible level of payoff: since directly maximising one's material payoff may lead to self-defeating behaviours, we suggested that rational players should be able to build *strategic* preferences, such that their satisfaction leads *in fine* to a higher level of material payoff. Our theory of preference formation therefore relies on the result that players can generally benefit from strategic commitment in games: we then identified the optimal weights each players should assign to the others in her utility function, and in particular the conditions under which this process could lead to the formation

of cooperative or competitive preferences. We showed that there exists a strong connection between the supermodularity of the initial game and the emergence of cooperative preferences. This result allowed us to discuss possible applications of our framework in terms of policy design: the efficient design of public policies should take into account the possible change in individual preferences induced by the policy. We then argued that public policies should privilege incentives that create a coordination game between the players and cut the possible submodularity of the initial game (as in climate change negotiations): this type of approach may indeed facilitate the formation of cooperative preferences, and hence facilitate the achievement of the policy objective.

Chapter 6

# The Rationale of Team Reasoning

## Contents




**Abstract:** Team reasoning theory has been developed to overcome the predictive failures of standard game theory in cooperation and coordination games: it is assumed that in certain situations, the individuals can conceive themselves as the members of a team, and make their choices so as to satisfy the team's objective. Unlike social preferences approaches in which we alter individual preferences, team reasoning implies a transfer of agency from the individual to the collective level, without requiring a modification of individual preferences. We develop a formal framework of team reasoning based on Bacharach (1999)'s notion of unreliable team

interaction, and define the collective preferences of the group as the result of a strategic choice. We show in particular that the construction of collective preferences may lead to aggressive behaviours toward the players outside the team in submodular games. We can then show that, in a very large class of games, team reasoning is the only procedure of choice that rational individuals can adopt.

## 6.1 Introduction

We have shown in the previous chapter that it is generally in the interest of rational players to choose preferences different from their own true preferences. In this chapter, we develop a model of preference formation grounded on Bacharach's variable frame theory and his model of team reasoning (Bacharach, 1999). Our model of preferences is consistent with our critique of the welfarist claim of LP, since it does not require the existence of true preferences (whose satisfaction would be normatively desirable), and probably offers a more realistic account of strategic interactions as relationships between agents, able to choose collectively what their own preferences are. The main contribution of our model is that it explicitly integrates the possibility of collective intentions. So as to illustrate the interest of this notion, imagine:

> **The conference in São Paulo**: we intend to go to a conference in São Paulo. We meet on Monday and agree on the most convenient flight and hotel. The plane tickets and the hotel cost the same, so we decide that you will buy the tickets while I will pay for the accommodation. On Tuesday, we are in our respective offices, I book the hotel while you independently buy the tickets. Later that day, we inform each other that each of us completed one's task. But was it actually rational for me to book the hotel?

Although the natural answer seems to be 'yes' (we indeed agreed on a joint action resulting from our collective intention of attending the conference), this is not so obvious from the perspective of game theory. Suppose for instance that we are both game theorists, and therefore that our rationality is common knowledge among us. As a rational individual, I should book the hotel if and only if I have a good reason to do so. In particular, since I have no reason to book a room in a city in which I will not be able to go, I will not book the hotel if I believe that you are not buying the tickets. I however know that you are rational too: you will therefore buy the tickets if and only if you have a good reason to do so. Since you do not want to arrive in São Paulo without accommodation, you will book the flight if and

only if you believe I booked a room for you. But you also have a reason to believe that I will not book a room for you: you may believe that I believe that you will not buy my ticket (and therefore, as a rational agent, I will not book the hotel), and this belief is possible if I believe you believe it (which is also a rationalisable belief if you believe I believe you believe it, and so on *ad infinitum*). It is therefore rational for me to book the hotel if and only if I believe that you believe that I believe... that one of us either buy the ticket or book the hotel. It is however also perfectly rational for me not to book the hotel: I can also believe that you believe... that one of us will either not buy the ticket or not book the hotel. My reason for booking the hotel was however probably much simpler: since we decided *together* to perform a specific course of action, I did not need another reason for booking the hotel than having been involved in the collective decision of sharing the task of booking the hotel and the flight. In particular, I did not need to justify in terms of your individual intention of attending the conference whether it was in your interest or not to buy the tickets: it was *our* intention of attending the conference that justified my choice, not the addition of yours and mine.

The idea that individuals can conceive themselves as members of a team — for whom collective instructions such as 'you buy the tickets and I book the hotel' constitute sufficient reason for action — has been developed in different forms by philosophers (Hodgson, 1967, Regan, 1980, Gilbert, 1989, Hurley, 1989, Hollis, 1998) and is closely related to the literature of collective intentions (Tuomela and Miller, 1988, Searle, 1990, Bratman, 1993). Such theories of 'team reasoning'[1] have then been introduced within economics by Sugden (1993, 2000, 2003) and Bacharach (1995, 1997, 1999, 2006). Several experimental studies suggest that team reasoning can explain how people can coordinate on focal points in coordination games (Bacharach and Bernasconi, 1997, Bardsley et al., 2006), but also that team reasoning is likely to occur in a broader class of games, including typically cooperation games (see for instance Guala et al. (2013) for a literature review).

Although theories of team agency may offer interesting perspectives for the study of cooperation and coordination games — in which a purely individual account of reasoning may lead to puzzling predictions — little work has been done to provide a complete game-theoretical framework to deal with collective agency. Sugden (1995), Casajus (2000), Janssen (2001, 2006) for instance use framing and team reasoning so as to provide a rationale for the selection of focal points. Zizzo (2004), Zizzo and Tan (2003) also suggest a 'game harmony' measure to capture

[1] Several names are used to refer to this idea of collective agency, such as 'we-thinking', 'team-thinking' or 'team reasoning'. We will use in this chapter the notion of 'team reasoning' since it is the label under which it has been introduced in game theory.

the degree of cooperation and conflict in games, and use it as an indicator of the likelihood of team reasoning. The main attempt to provide such a formalisation of team reasoning is however due to Bacharach (1999, 2006). Bacharach indeed intended to explain the emergence of team reasoning thanks to game theory, i.e. to model *how* people team reason, but also *why* they are likely to team reason. He proposes a formal theory of games in which players can be either I-reasoners or we-reasoners (Bacharach, 1999). The mode of reasoning of the players is determined by psychological phenomena, prior to rational choice. The individual has a specific perception of the game — the *frame* through which she perceives the game —, either as an 'individual' or as a 'member of a group', which is exogenously given, and then chooses the optimal action either for herself or for the group. If she is a 'we-reasoner', she computes the optimal strategy from the perspective of the group, and then plays her part of this joint profile. As stated in his 1999's paper, Bacharach then mainly provided a theory explaining how people may team reason ('given that someone team reasons [...], to what choice does this lead her? (Bacharach, 1999, 142)), and left for future work the complementary question of whether it is rational or not to team reason. He unfortunately unexpectedly died before being able to complete his work[2]. Several justifications of team reasoning are sketched in his 2006's book, relying on the notion of game harmony (developed later by Zizzo (2004), Zizzo and Tan (2003)) and the interdependence hypothesis (studied by Smerilli (2012)), but his main argument was on an evolutionary level, by appealing to multilevel selection theory. Unlike the standard interpretation of evolution theory, this theory assumes that natural selection also operates at the level of the groups. Although team reasoners have a lower fitness than individualistic reasoners, the groups to which they belong are likely to outperform groups with only a little fraction of team reasoners. It is therefore possible that team reasoning remains an evolutionary viable trait, if the diminution of the share of team reasoners within each group is outweighed by the increase in size of the groups with a high share of team reasoners (this is a typical case of the Yule-Simpson effect).

The aim of this chapter is to provide a game-theoretic framework for team reasoning, so as to treat the two main issues suggested by Bacharach — *how* and *why* do people team reason. We will keep the main features of the model of team reasoning developed in the 1999's paper, such as the 'framing' perspective of collective intentions, while offering a slightly different definition of an 'unreliable

[2]His 2006's book was published posthumously, edited by Gold and Sugden. They collected the materials Bacharach intended to insert in the book and discuss Bacharach's plans for his uncompleted chapters.

team interaction' — Bacharach's notion of games in which people are able to team reason. Furthermore, and unlike the different works cited above, we will investigate further the determination of the collective preferences of the team — the preferences that we, as a group, want to satisfy — by suggesting a model of strategic choice of collective preferences. We will in particular study how groups should interact with other groups, i.e. whether they should cooperate or compete with the others. Our analysis highlights that team reasoning may be developed within a standard game theoretic framework, by using appropriate notions of preferences and equilibrium. More specifically, we show that the equilibrium of an unreliable team interaction is equivalent to a specific refinement of Bayesian Nash equilibrium in a game of incomplete information in which types define a specific class of collective preferences. We then study the conditions under which it is rational to team reason, and show that, under the assumption of common knowledge of rationality, it is generally in the interest of the players to team reason.

The rest of this chapter is organised as follows. In section 6.2, we defend the relevance of introducing a theory of collective intentions in game theory. Section 6.3 describes the basic framework of variable frame theory and present Bacharach's model of team reasoning. We then characterise in section 6.4 the optimal collective preferences team reasoners should satisfy, before determining in section 6.5 the conditions under which it is actually rational for them to team reason. Section 6.6 concludes.

## 6.2 Collective intentions and social preferences

In this section, we argue in favour of team reasoning instead of social preferences approaches as the basis of an economic theory of unselfish behaviour. We firstly present some theoretical difficulties of standard game theory in cooperation and coordination games (subsection 6.2.1). We then discuss the traditional solution consisting in modifying individual preferences by integrating social preferences in individual utility functions. We highlight several difficulties of this approach, such as its ambiguous interpretation and its inability to properly solve coordination games (subsection 6.2.2). We finally argue that team reasoning provides a more satisfying solution to solve cooperation and coordination puzzles (subsection 6.2.3).

### 6.2.1 Two puzzles of game theory

The common representation of human behaviour in economics is based on two fundamental assumptions, rationality and self-interest: individuals are assumed to act rationally in pursuit of their objectives, and these objectives are assumed to be defined in terms of their own well-being (Sen, 2002). Understood as a normative theory of behaviour, the 'self-interest theory'[3] states that an individual should choose the outcomes that would be best for herself, according to her beliefs about the state of the world. This theory of individual choice has then been extended to model behaviour in strategic interaction, i.e. a choice under uncertainty in which the state of the world depends on the choices of other rational individuals. Aumann (1987) for instance suggests that the choice of others is part of the description of the state of the world, and show that individual Bayesian rationality (i.e. the maximisation of one's expected utility given one's beliefs) leads to a correlated equilibrium: solution concepts in games may therefore be seen as the *result* of standard Bayesian rationality[4] (the adequate concept would then depend on the assumptions we make on the prior beliefs (Bernheim, 1986)).
This connection has been stressed by Bacharach and Hurley (1991), who argue that players' choices in games are derived from expected utility theory, i.e. a 'theory of rational individual choice under uncertainty', as well as Binmore (1994), according to whom:

> Game theorists of the strict school believe that their prescriptions for rational play in games can be deduced, in principle, from one-person rationality considerations without the need to invent collective rationality criteria —provided that sufficient information is assumed to be common knowledge. (Binmore, 1994, p.142)

The object of game theory is then to 'propose *solutions* for games' (Sugden, 2001, p.115, emphasis in original), a solution being defined as a profile of individual strategies that respects some desirable properties. Those properties define a solution concept, which select for each game within a class of games (e.g. normal form games) a strategy profile — the solution of the game. It is commonly agreed that a solution concept must (i) be consistent with the rationality of the players,

[3] We use here the terminology of Parfit (1984).

[4] Mariotti (1995) nevertheless highlights that two of Savage (1954)'s axioms of the Bayesian model — ordering and dominance — are incompatible with some basic game theoretic principles: this suggests that modelling the decision of others as part of the description of the state of the world may lead to an unsatisfactory theory of strategic interaction.

and (ii) assign a restricted number of solutions per game (ideally a single one for each game) (Sugden, 2001, p.115). A solution concept determines the strategy a rational player should play in a game (from a normative perspective), and may therefore also provide some insights about how real individuals actually choose, if we assume that real individuals have a tendency to act rationally. A common assumption of game theory is then to assume that individual rationality is common knowledge, such that the players may anticipate the behaviour of the other rational agents. The most widely used solution concept is Nash equilibrium, since it precisely embodies this condition of individual rationality (no player should have an interest in unilaterally deviating from her strategy, knowing that the others play their equilibrium strategies), and its existence has been shown in a wide class of games of interest for economists (Nash, 1950, Debreu, 1952). Assuming that the players are rational and that this rationality is common knowledge — as assumed in standard game theory — however leads to puzzling predictions in two families of games, *cooperation* and *coordination* games.

The most famous illustration of a cooperation game is the prisoner's dilemma (PD):

| | C | D |
|---|---|---|
| C | (R;R) | (S;T) |
| D | (T;S) | **(P;P)** |

with $T > R > P > S$. If we assume that players choose their strategy so as to maximise their expected utility, given their beliefs about the strategy of the other, then a rational player must necessarily play $D$ (it is indeed a strictly dominant strategy). As a normative theory of choice, the self-interest theory recommends each player to choose $D$. As a descriptive theory of choice, it says that we should observe a significantly higher proportion of individuals playing $D$. The puzzle here is that experimental evidence highlights that real individuals actually cooperate in unrepeated prisoner's dilemma almost half of the time (Sally, 1995). It seems therefore that the self-interest theory does not provide an adequate descriptive theory of choice. More problematically, it appears that, although it is individually rational to play $D$, reaching the equilibrium $\{D; D\}$ is not collectively rational: it would be better for both players to reach the profile $\{C; C\}$ instead of $\{D; D\}$. This means that following the recommendations of the self-interest theory will lead you *in fine* to a worse situation — in terms of your objective — than if you had followed the recommendation of another theory. For instance, rational individuals

who intend to maximise their expected utility in a prisoner's dilemma will achieve a lower outcome that if they were both trying to maximise the payoff of the other player. The normative failure of the self-interest theory results from the fact that this theory is indirectly self-defeating (Parfit, 1984), e.g. that 'irrational' individuals (as the one found in lab experiments) can actually achieve more than rational individuals.

The second puzzle of game theory is the issue of equilibrium selection in coordination games. In those games, the interests of the players are perfectly aligned and lead to an issue of equilibrium indeterminacy:

| | A | B |
|---|---|---|
| A | **(R;R)** | (0;0) |
| B | (0;0) | **(P;P)** |

with $R \geq P > 0$. If $R = P$, the game is a pure-coordination game, and game theory cannot predict which equilibrium the players should choose. Several empirical studies however show that real individuals successfully exploit apparently irrelevant features of the choice environment to determine their choice (Schelling, 1960, Mehta et al., 1994a,b, Bacharach and Bernasconi, 1997), such as the lexicographical order of the strategies in the above example ('A' indeed appears to be more salient than 'B', and actual players may therefore be tempted to play A because they expect the others to recognise A as a salient strategy too).
More strikingly, if $R > P$, we are in presence of a Hi-Lo game, in which one equilibrium Pareto dominates the other one. Although the choice to make in such a situation seems obviously trivial, standard game theory fails to predict why almost everyone choose A (Bacharach, 2006, 43). The conceptual issue here is that standard game theory assumes that individuals are instrumentally rational given their beliefs, i.e. the only reason for action that the players recognize is maximizing their expected utility: P1's choice therefore depends on what P2 is expected to do; P1 will therefore play A if and only if P2 is more likely to play A. But P1 knows that P2 wants to maximize her expected utility too, and P2's choice depends on what P1 is expected to do. A difficulty arises because there is no reason for P1 not to play B — since B will be the best reply of P1 if she believes that P2 will play B, this belief being possible by the fact that P1 may play B if she believes that P2 will also play B. Although the Hi-Lo game is not a puzzle at all for real individuals, it appears that rational players in the sense of game theory are unable to unambiguously select the dominant equilibrium.

The different approaches developed in order to explain this behavioural fact consist generally in restricting individual rationality, by assuming for instance magical thinking (Elster, 1989, 195-202) or that $\{A; A\}$ is the most salient equilibrium (Harsanyi and Selten, 1988) — but without providing a rationale for the selection of salient equilibria. Although those approaches explain the behavioural fact that people tend to play A, they do not provide a theory explaining why A obviously seems to be the right (and rational) choice. Sugden (2001) argues that this issue of equilibrium selection contributed to an important shift in game theory from models based on assumptions of ideal rationality towards evolutionary models, since the selection of a specific equilibrium is explained by path-dependencies and historical contingencies. Kandori et al. (1993) for instance consider an evolutionary model with a finite number of players and random mutations, and define the solution concept of 'long run equilibrium' as the limit distribution of a nonlinear stochastic difference equation. They show that, for symmetric $2 \times 2$ games with two symmetric strict Nash equilibria, the long run equilibrium satisfies Harsanyi and Selten (1988)'s criterion of risk dominance, and corresponds to the dominant equilibrium if the strategies have equal security levels. This means that $\{A; A\}$ is the unique long run equilibrium in a Hi-Lo game. However, although their solution concept provides a rationale for selecting the dominant equilibrium in a Hi-Lo game, it also implies that rational players should coordinate on the Pareto dominated equilibrium in a Stag Hunt (which is the risk-dominant equilibrium). Similarly to the prisoner's dilemma, it seems normatively questionable not to select the payoff dominant equilibrium (this solution concept indeed leads to a self-defeating behaviour) and empirical results suggest that, although players are quite sensitive to change in the risk dominance characteristics of the game (Schmidt et al., 2003), many factors related to the structure of payoffs may have a significant role on the equilibrium selection. Battalio et al. (2001) and Dubois et al. (2012) for instance suggest that the 'optimisation premium' — the expected earnings difference between the two actions — plays a crucial role in the process of equilibrium selection (more than the best-response correspondence), and that real individuals tend to play the strategy of the payoff dominant equilibrium when this premium is relatively low. Using the long run equilibrium as a predictor for actual choices may therefore lead to unsatisfactory results, since only the risk-dominant equilibrium can be selected.

### 6.2.2 Individual rationality with collective preferences

The main approach suggested so as to deal with the empirical evidence of unselfish behaviours in cooperation games has been to relax the assumption of self-interested

preferences: Becker (1996, p.4) assumes for instance that 'individuals behave so as to maximize utility while extending the definition of individual preferences to include [...] love and sympathy, and other neglected behavior'. The existence of other-regarding preferences can for instance be justified as the expression of a utilitarian ethics: a utilitarian should indeed always choose her action so as to promote the greatest good for the greatest number. Edgeworth (1881) suggest for instance that there may exist 'mixed modes of utilitarianism' (pp.102-104), and that we can deduce the utility function of the individual from her material payoff 'by multiplying each pleasure, except the pleasures of the agent himself, by a fraction – a factor doubtless diminishing with what may be called the social distance between the individual agent and those of whose pleasures he takes account' (p.103). Since the utility of such an 'altruistic' individual is derived both from her material payoff and (positively) from the material payoff of others, sufficiently altruistic players may spontaneously cooperate with other individuals in a PD — it is indeed in their own interest to do so. The possibility of cooperation is therefore not due to a wrong theory of choice, but to a misrepresentation of the game: since individuals' utilities do not correspond to their material payoff, it means that the game they were facing was not a PD, but another game in which full cooperation is a Nash equilibrium.
Another and more recent variant of this social preferences approach has been suggested by Fehr and Schmidt (1999) (and then studied by many others, such as Bolton and Ockenfels (2000) or Charness and Rabin (2002)), who assume that individuals are averse to inequality in the distribution of payoffs. This approach however remains remarkably similar to the previous explanation in terms of altruism. It indeed relies on the central idea that, if standard theory fails to predict the right outcome, then it means that the utility functions were ill-specified. The core of all these theories of unselfish behaviour is not that people are altruistic in the sense that they directly care about other's payoff: they are selfish — in the sense of psychological egoism — but their utility is increased when promoting other's well-being. In particular, the second core assumption of individual rationality remains central. Bicchieri (2004, p.183) describes this approach as follows:

> The theory of rational choice's central assumption is that a decision maker chooses the best action available according to her preferences. The content of preferences is unrestricted. Agent's preferences may be selfish or altruistic, self-defeating or even masochistic. Preferences mirror values and dispositions that are beyond the pale of rationality. What is required is that preferences are well behaved in the sense of fulfilling certain formal conditions [...]. If preferences are well behaved, they can be represented

> by utility functions, and rationality consists in maximizing one's utility function, or finding the maximum value of one's utility function.

A difficulty of this kind of approach is the interpretation we should give to such 'social preferences': it is indeed unclear whether those preferences are immutable, i.e. whether cooperating in a social dilemma truly reflects empathetic concerns for other humans (such as inequity aversion) or simply reflects the compliance to what the individual think is the social norm in the game (see for instance Gintis (2009), Binmore (2010) and the discussion between Binmore and Shaked (2010), Eckel and Gintis (2010) and Fehr and Schmidt (2010)). For instance, if we assume that cooperative behaviours observed in experiments highlight the existence of a concern for fairness, and that individuals are instrumentally rational given their social preferences, then two players, both averse to inequity, should be averse to inequity not uniquely in social dilemmas. Their concern should remain stable for other games. Consider for instance the two following games:

| **G1** | C | D |
|---|---|---|
| C | $(3;3)$ | $(0;4)$ |
| D | $(4;0)$ | **(2;2)** |

| **G2** | C | D |
|---|---|---|
| C | **(3;3)** | $(4;0)$ |
| D | $(0;4)$ | $(2;2)$ |

The game $G1$ is a simple prisoner's dilemma, whose unique Nash equilibrium $\{D;D\}$ is Pareto dominated by the cooperative profile $\{C;C\}$. The second game $G2$ is a similar situation in which the payoffs have been inverted. In this case, there is still a unique pure strategy Nash equilibrium $\{C;C\}$, but it is the Pareto optimal outcome. Suppose now that P1 and P2 are averse to inequity, such that their utility functions — as in Fehr and Schmidt (1999) — are as follows:

$$\begin{aligned} U_1(x;y) = \Pi_1(x;y) &- \alpha_1 \max(\Pi_2(x;y) - \Pi_1(x;y);0) \\ &- \beta_1 \max(\Pi_1(x;y) - \Pi_2(x;y);0), \end{aligned} \tag{6.1}$$

with $\Pi_i(x;y)$ the material payoff (as described in the payoff matrix) of individual $i$, $0 \leq \beta_i < 1$ and $\alpha_i \geq \beta_i$. Consider that $\alpha_1 = \alpha_2 = 1$ and $\beta_1 = \beta_2 = 2/3$. The utilities — determining the choice of the players — for those two games are the following:

| **G1'** | C | D |
|---|---|---|
| C | **(3;3)** | $(-4; 4/3)$ |
| D | $(4/3; -4)$ | **(2;2)** |

| **G2'** | C | D |
|---|---|---|
| C | **(3;3)** | $(4/3; -4)$ |
| D | $(-4; 4/3)$ | **(2;2)** |

In the transformed PD, $\{C; C\}$ constitutes now a second Nash equilibrium, explaining the coexistence of cooperative and non cooperative behaviours in social dilemmas. However, although inequity aversion seems to offer a relevant theory of unselfish behaviour, it also implies a quite odd result in the second game $G2'$: we find indeed that $\{D; D\}$ is now also a Nash equilibrium, which is here Pareto dominated. If the players are concerned by inequity, then we should also find empirical evidence of individuals playing the strategy D in $G2$, although this strategy is strictly dominated and lead to a suboptimal equilibrium. If we accept inequity aversion as the explanation of cooperation in the PD, we should also expect finding puzzling predictions in games like $G2$.
Explaining unselfish behaviours by altering individual preferences should also be consistent with the more trivial games in which self-interested players reach an optimal equilibrium. If we consider that individual preferences may change according to the payoff structure of the game faced by the individual (in the sense that the players are inequity averse in $G1$ but not in $G2$), then the theory is tautological and lose its predictive interest — since we will define *post hoc* preferences.

Therefore, if we assume that individual preferences remain identical over the two games, assuming the existence of social preferences is probably not a satisfying approach to explain why the players can cooperate in both games and defect only in the first one: for given preferences, they must adopt a different approach for each game. It seems therefore that a more general theory of unselfish behaviour should encapsulate this idea that the individuals do not interact with others always in the same fashion and, for given preferences, may present different *intentions*. We can find this idea with Rabin (1993), who makes the assumption that players take into account within their decision process other players' intentions: if P1's action increases P2's payoff at some cost for P1, then P1 is categorised (from P2's perspective) as 'kind'. Such 'kindness' is the evidence that P1 derives utility from P2's payoff, and the behavioural assumption that follows is that people tend to

be kind with the individuals who were kind to them (and conversely, people are unkind with people who are unkind to them). Rabin however recognizes that his theory may lead to unsatisfactory results in sequential games (in particular in the trust game, since trusting the second player cannot be interpreted as a sign of kindness or unkindness: P1 indeed chose to trust P2 because she believed P2 would honour her trust; it can therefore be seen as a perfectly self-interested behaviour) and requires to model 'additional emotions' (Rabin, 1993, 1296). More generally, the literature on psychological games (Geanakoplos et al., 1989, Dufwenberg and Kirchsteiger, 2004, Falk and Fischbacher, 2006, Falk et al., 2008) rests on the hypothesis that utility functions can depend on both actions and beliefs about those actions: although a clear distinction is made between the material payoff of the players and their utility, this approach can be seen in the same revisionist perspective than the one described above. Colman (2003) argues that this approach is relatively limited to the extent that it does not offer a general framework to explain cooperation puzzles such as cooperation in the Centipede game (see the discussion between Colman (2003) and Carpenter and Matthews (2003)): explaining cooperative behaviours thanks to individual intentions probably requires to directly question the formal principles of reasoning of the individuals. This is precisely what we intend to do in this chapter by developing a formal framework of team reasoning[5].

The last (and probably main) difficulty of keeping the assumption of individual rationality while extending the set of definition of individual preferences is that this approach is unable to solve coordination puzzles like the Hi-Lo game:

| | A | B |
|---|---|---|
| A | **(2;2)** | $(0;0)$ |
| B | $(0;0)$ | **(1;1)** |

In this kind of game, assuming the existence of social preferences is totally ineffective, unless we define arbitrary utility functions such that the unique Nash equilibrium will be $\{A;A\}$. Such payoff distortions cannot however easily be justified as the expression of a specific feeling or disposition towards the other players: they are just the *ad hoc* distortion that enables the modeller to explain an empirical regularity. The difficulty of coordination games is indeed not linked to the payoff matrix and individual incentives (on which social preferences approaches may have

[5]In addition to team reasoning, Colman (2003, pp.149-152) also suggests Stackelberg reasoning (Colman and Bacharach, 1997) and non-monotonic reasoning as alternative to the standard game-theoretic mode of reasoning.

an influence), but to the mode of reasoning of the players, and it seems that the only way to solve coordination games is to assume that players can adopt other modes of reasoning than Bayesian rationality. This claim is defended for instance by Bardsley (2007), who argues that solving coordination games requires to introduce a non-reductive account of collective intention (i.e. a notion of intention different from the individual rationality postulated in game theory).

### 6.2.3 Collective rationality with individual preferences

An alternative approach for explaining non-selfish behaviours in games would be to question the assumption of individual rationality, by suggesting the possibility of collective agency. The schema of practical reasoning (Gold and Sugden, 2007) of a *I-reasoner* — an individually instrumental rational individual, in the sense of standard game theory — can be described as follows:

(1) I must choose between C and D;

(2) if I choose C, the outcome will be $u(C)$;

(3) if I choose D, the outcome will be $u(D)$;

(4) I prefer achieving $u(D)$ than achieving $u(C)$;

- I should choose D.

This kind of reasoning implies that, for given beliefs about the strategy of other players, a rational individual should play her individual best reply. In the case of a prisoner's dilemma, rational players should always choose D. Consider now the schema of a collective instrumental reasoning:

(1) We must choose between $\{C;C\}$, $\{C;D\}$, $\{D;C\}$ and $\{D;D\}$;

(2) if we choose $\{C;C\}$, the outcome will be $(R;R)$;

(3) if we choose $\{D;D\}$, the outcome will be $(P;P)$;

(4) if we choose $\{C;D\}$, the outcome will be $(S;T)$;

(5) if we choose $\{D;C\}$, the outcome will be $(T;S)$);

(6) we prefer achieving $(R;R)$ than achieving $(P;P)$;

- we should not choose $\{D;D\}$.

Gold and Sugden (2007) argue that, due to the symmetries between the two kinds of reasoning, we cannot accept one reasoning without accepting the other. Since the conclusion of those two valid reasonings are in contradiction (two individually rational players should play $\{D; D\}$, while two collectively rational players should not), their premisses must be incompatible: in the case of individualistic reasoning, the unit of agency is 'I', whose motivation is achieving *my* preferred outcome, while the collective reasoning procedure relies on the fact that 'we' is the unit of agency, whose motivation is achieving *our* preferred outcome. Although 'I' prefer to defect and 'you' prefer to defect, 'we' prefer to cooperate.
Our collective intention to maximise our individual payoffs is therefore not reducible to the sum of our individual intentions to maximise our individual payoffs. The possibility of collective agency implies that individual may conceive themselves not as 'individuals' but as 'members of a team'. The mode of reasoning is radically different: a *team reasoner* is not pursuing her personal objectives, but the objectives of her group. From an individual perspective, the basic features of the reasoning of a team reasoner may be described as follows:

(1) we are the members of a group S;

(2) each of us identifies with S;

(3) we must choose between $\{C; C\}$, $\{C; D\}$, $\{D; C\}$ and $\{D; D\}$;

(4) we prefer achieving $(R; R)$ than achieving $(P; P)$, $(T; S)$ or $(S; T)$;

(5) each of us wants that we choose what we prefer;

- each of us should choose 'C'.

The central idea of this schema is that each of us is doing her part in the joint action resulting from our collective intention to satisfy our preferences. A team reasoner is pursuing the objective of the team (the satisfaction of the collective preferences), and chooses her strategy as if she were the member of a coalition — in which an imaginary supervisor attributes to each individual her 'part' of the collective strategy profile:

> Choosing as a member of a team entails not only being motivated by the team objective, but also a different pattern of reasoning: an agent who 'team reasons' computes, and chooses her component in, a profile evaluated using the team's objective function (Bacharach, 1999).

The crucial difference with social preferences approaches — whose explanation of cooperation relies on other-regarding preferences, and therefore only question the assumption of non-tuism — is that team reasoning is neither selfish nor altruistic: it represents individuals as reasoning together about the achievement of common goals (Sugden, 2011). It is therefore not the preferences of the individuals that are ill-specified, but their mode of reasoning. This explanation of cooperation may be more relevant than social preferences approaches: we can indeed question the relevance of grounding a theory of cooperation on altruism, since altruism does not require reciprocity. If P1 cooperates in a prisoner's dilemma, it is probably not because she wants to maximize the payoff of P2, but because she expects P2 to cooperate too (so that they will be able to achieve *together* the 'good' outcome). Cooperation is not a matter of altruism (which is one-sided): it is the process of a group of individuals working together so as to promote a mutual advantage. Although we do not question the existence of altruistic motives or concerns for inequity aversion, we would like to argue that cooperation is probably more a question of reciprocity and teamwork rather than of altruism.

As an illustration, consider the two games G1 and G2 discussed above:

| **G1** | C | D |
|---|---|---|
| C | (3; 3) | (0; 4) |
| D | (4; 0) | (2; 2) |

| **G2** | C | D |
|---|---|---|
| C | (3; 3) | (4; 0) |
| D | (0; 4) | (2; 2) |

Suppose that P1 and P2 are team reasoners, and that this is common knowledge. Each player therefore perceives the games as the choice of 'us' (henceforth $N$) against nature. The games can therefore be represented as follows:

| **G1'** | CC | CD | DC | DD |
|---|---|---|---|---|
| N | (3;3) | (0;4) | (4;0) | (2;2) |

| **G2'** | CC | CD | DC | DD |
|---|---|---|---|---|
| N | (3;3) | (4;0) | (0;4) | (2;2) |

Suppose also for convenience that the preferences of $N$ are such that the vector of payoffs (3;3) is collectively preferred to (2;2), (0;4) and (4;0). We can easily

justify that (3;3) is preferred to (2;2) by an argument of Pareto dominance, while the preference of (3;3) over (4;0) and (0;4) can be justified either by appealing to a utilitarian principle (as maximising the sum of payoffs) or a principle of fairness. In those two games, the optimal choice for $N$ is $\{C;C\}$. Given the optimal strategy profile, each player therefore plays her part of this profile, C. However, when mutual membership in $N$ is not certain — $N$ is therefore an *unreliable team* — it is not certain that a team reasoner will necessarily play $C$. Indeed, if P1 is a team reasoner but believes that there is a high chance that P2 is a I-reasoner, then P1 is likely to defect, since she expects P2 not to do her part of the collectively rational strategy profile $\{C;C\}$.
Team reasoning therefore offers a relatively simple explanation of cooperative behaviours in prisoner's dilemma, but also — and unlike psychological games — in sequential games like the Trust or the Centipede games. We can indeed notice that, since the optimal choice is computed by the team, the paradoxes associated to backward induction disappear: $N$ chooses (alone) the optimal path of strategies, and the players of the team are committed to respect this path (see for instance Hollis (1998, p.137) and his analysis of the 'Enlightenment trail').

Furthermore, although it is possible to explain prosocial behaviours by appealing either to social preferences or to collective agency, we showed that modifying the content of individual preferences is insufficient to solve coordination games, whereas team reasoning can tackle this kind of issues. In the Hi-Lo game, 'we' (as a unit of agency) wants to obtain the Pareto-dominant payoff $(R;R)$: as a member of the team 'we', I shall therefore play A. The core argument of team reasoning is that 'players who think as a team do not need to form expectations about one's another's actions' (Sugden, 1993, 87), and this is why they are able to avoid the infinite regress faced by I-reasoners: unlike standard game theory in which players treat the actions of others as fixed (and are then maximising their expected utility given the strategy of others) the expectations of the team reasoners are about mutual team membership and not actions. Sugden (1993) then argues that, if it is common knowledge that we (P1 and P2) are both members of the unit 'we' (the team), then it is rational for us (P1 and P2) to act according to the prescriptions of team reasoning, i.e. to follow the instructions of 'we'.

We have shown that only team reasoning is able to solve simultaneously co-operation and coordination puzzles: this implies that team reasoning may offer a better foundation for a parsimonious theory of unselfish behaviours, rather than the

revisionist strategy consisting in altering individual preferences while keeping an individualistic account of agency. It must also be noticed that explaining cooperation and coordination thanks to team reasoning is less demanding from an epistemic perspective than their standard explanation with Bayesian rational players. Indeed, while standard game theory requires (i) beliefs on the rationality of the other players, and (ii) beliefs on their actions, team reasoning only requires beliefs on the rationality of the other players. Bayesian players indeed choose the strategy that maximises their expected utility given their prior beliefs on the *action* of the other players, and then adapt those beliefs according to their beliefs on the *rationality* of the others — i.e. whether the others are rational too, and whether this is known or not. Team reasoning on the contrary only requires the players to have beliefs about the rationality of the others (whether they are I-reasoner or team-reasoners, and whether this is known or not), not about their actions. If it is common knowledge that all the players are team reasoners[6], then they do not need to form beliefs about the action of the others: they are indeed choosing the collective profile from the perspective of the team, and are then committed to play their part of this profile — knowing that the others are team reasoners is a sufficient reason to ensure them that the others will effectively play their part of the collective profile.

## 6.3 Team reasoning and frames

We present in this section Bacharach's attempt to model team reasoning, and develop our own model based on it. We firstly describe the basic framework of Bacharach's variable frame theory (section 6.3.1), on which he grounds his model of team reasoning. We then present our model of team reasoning based on Bacharach's model (section 6.3.2) and study the connections between our model and Bayesian games (section 6.3.3).

### 6.3.1 Variable frame theory

Variable frame theory (Bacharach, 1991, 1993, 1999, 2006) provides an analysis of games that explicitly takes into account players' framing of the decision problem (see Larrouy (2013) for a detailed presentation). The primitive of the analysis is the existence of an *objective* choice problem, defined in terms of players, space of

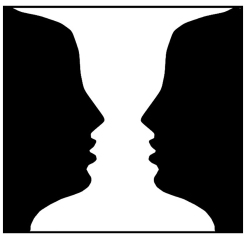
[6]Note however that the common knowledge of team reasoning is not required to ensure that some players will team reason: we will indeed develop our model simply by assuming that the players have a common prior belief on the mode of reasoning of each player. This assumption 'replaces' the standard assumption of common knowledge of rationality.

strategies and states of the world. The crucial departure from standard game theory is that the decision problem faced by each individual is not this objective problem: taking a decision indeed requires for each player to build a subjective representation of the situation of choice. A player's frame therefore corresponds to the set of variables she uses to conceptualise the game (Bacharach, 1997, 4). Once the choice problem has been framed, the individual evaluates the different alternatives with respect to the way she framed it. The act of choosing is therefore a two step process: (i) the framing of the objective game (which is prior to reasoning), and (ii) the strategic choice of the individual given her representation of the game. The main difference with the standard literature on framing effects (Kahneman and Tversky, 2000) is that there is no objectively identifiable set of choice. A choice set is meaningful only if it has been framed in a certain way. Those frames are therefore not biases against some objectively unique way of looking at the choice problem. This implies that individual preferences are not defined with respect to the objective game, but with respect to each possible framing of the game.
Variable frame theory was thought as a way to 'outline a rigorous theory of salience' (Bacharach and Bernasconi, 1997, 2) so as to solve coordination puzzles. Bacharach indeed intended to provide a theory within which the act of choosing a salient option should be the result of a rational choice, and not simply due to the mere fact that this specific option is actually salient (Bacharach and Bernasconi, 1997, 34). In this chapter, we will however not deal with issues of equilibrium selection in pure coordination games, and restrict our analysis of variable frame theory as the support of collective agency and team reasoning[7]. Our main objective here is indeed to explain the possible emergence of team reasoning in non cooperative games.

Bacharach (2006) suggests two kinds of frames so as to justify team reasoning from the perspective of variable frame theory, the I-frame and we-frame. Standard game theory implicitly assumes that the players perceive the game trough the I-frame, and therefore ask themselves 'what do I want, and what should I do to achieve it?'. Players perceiving the game through a we-frame will however ask themselves 'what do we want, and what should I do to help achieve it?'. While players with a I-frame think that they are playing the game *against* the others, players with a we-frame think that they are playing the game *with* the others. Bacharach (1999) provides a formal theory modelling the strategic interactions of players presenting I- and we-frames, but was not able to complete his work and to

[7] On the selection of focal points in a formal game theoretic framework, see for instance the works of Casajus (2000) and Janssen (2001, 2006).

model the process of selection of frames. He informally suggests the 'interdependence hypothesis' as an explanation of the emergence of team reasoning. In a nutshell, this hypothesis states that, if an outcome that can be reached by an individually rational reasoning is Pareto dominated by an outcome that can only be reached by a collective reasoning (such that the cooperative outcome in a PD), then group identification and the selection of a we-frame is more likely (Smerilli, 2012). Bacharach (2006) also provided arguments to justify from an evolutionary perspective the selection of frames: while I-frames will be more successful at the individual level, we-frames will favour groups of players whose proportion of team reasoners is higher. This combination of individual and group selection (multilevel selection theory, see Sober and Wilson (1998) for a presentation) implies that, although the share of team reasoners tends to decrease within each group over time, the groups with a high share of team reasoners tend to grow faster, while non cooperative groups progressively disappear: it is therefore possible that the share of team reasoners in the total population eventually increases.
So as to illustrate Bacharach's notion of 'frame', consider the following Rubin vase:



There are two main representations of this object: we can either perceive two black faces on a white background, or a white vase on a black background. An interesting property of ambiguous images such as the Rubin vase is that we cannot perceive *simultaneously* the faces and the vase, although we may know that both perceptions are possible. This means in particular that there is no 'good' perception of the object: the two perceptions in terms of faces and vase are both equally valid perceptions. Within the framework of variable frame theory, it is assumed that there exists an object that is beyond human perception, which determines our possible perceptions of it[8]. The observer has then two main ways of framing this object: either as 'two faces' or as 'one vase'.

[8]We could for instance consider that a 'Rubin vase' is a graph, and the 'true object' would correspond to the equation describing this graph. As long as we do not represent the graph on a sheet of paper, we cannot perceive neither the vase nor the faces.

Consider now the case of games in normal form, and more specifically of a PD. The primitive of the game is a mathematical object $G = \langle N, X, \Pi \rangle$, with $N$ the set of players, $X$ the set of strategy profiles, and $\Pi$ the vectors of utilities associated to each strategy profile. We however represented this game in the previous section in a very specific way: we attributed to each player a label (P1 and P2), different labels for the two available strategies (C and D), and represented the game as a payoff matrix (we represented the different states of the world by the material payoff each player gets). Variable frame theory states that the players cannot choose directly from $G$, since the act of choosing requires a specific perception of the game. In particular, the game as represented by the theorist is not the objective game, and it is possible that the players do not perceive the game in the same way. For instance, while the game theorist perceives the game $G$ as $G1$, individuals averse to inequity perceive it as $G1'$, and team reasoners as $G1'$.

### 6.3.2 Unreliable team interactions

We now turn to a more formal presentation of variable frame theory and team reasoning. Our formalism is slightly different from Bacharach (1999), although it will share many of its features: our objective is to reformulate Bacharach's model of team reasoning so as to highlight the connections between unreliable team interactions and Bayesian games.

Let $N = \{1; \ldots; n\}$ denote the finite set of players, and $X_i$ the set of strategy of player $i \in N$. $-S$ denotes the complementary set of $S$ in $N$. Let $\mathfrak{S}$ denote the set of possible states of the world $\mathfrak{s}$ that may result from a decision $x \in X$. $\mathfrak{s}$ therefore describes the 'objective' and unframed state of the world: the preferences of player $i$ cannot be directly defined on $\mathfrak{s}$, but only on the perception $i$ has of the state of the world. In this chapter, we will consider only two kinds of frames, the 'I-frame' and the 'we-frame', i.e. the assessment of the state of the world when the individual conceives herself as an isolated individual or as a member of a group. The set of possible frames of $i$ therefore consists in $\mathcal{S}_i = \{S \subseteq N \mid i \in S\}$, i.e. the set of coalitions to which $i$ belongs. The distribution of frames across the players is denoted $\mathbf{S} = \{S_i\}_{i \in N}$.

An individual $i$ has therefore several ways of framing the game: the frame $S_i$ (her 'team') through which she will perceive the game is randomly chosen according to a distribution $\theta_i \in \Theta_i = \Delta(\mathcal{S}_i)$. An individual with the frame $S_i$ identifies with the coalition $S_i$, and act as if she were playing cooperatively with the members of her coalition. The resulting game is therefore a situation in which several coalitions are in competition, although the effective coalition structure $\mathbf{S}$ of the game is not

necessarily observed: when $i$ identifies with $S_i$, she cannot be sure that $j \in S_i$ also identifies with $S_i$. Although I may perceive a PD as a game of 'us' against nature (in which *we* should choose $\{C;C\}$), I am not sure that you share this perception. The teams $S_i$ are therefore *unreliable teams*.
We also assume that individual types (the individual probability distribution of frames $\theta_i$) are common knowledge. There therefore exists a common prior $\theta \in \Theta$ on the distribution of frames $\mathbf{S} \in \mathcal{S}$. Furthermore, we can derive from $\theta$ a belief $\mu_i : \mathcal{S}_i \mapsto \Delta(\mathcal{S}_{-i})$, corresponding to the distribution of frames player $i$ expects among the players $j \neq i$, knowing that her frame is $S_i$. We have therefore $\mu_i(S) = \theta_{-i}(S_{-i}|S_i = S)$.
We must therefore now define the preferences of the players under the different frames: while the preferences under the I-frame will be defined as in standard theory, we will investigate in section 6.4 what should be the preferences of the individual under a we-frame, i.e. when she is actuated by collective intentions and not only by the satisfaction of her own preferences.

We assume that the preferences $\succeq_i$ of player $i$ under the I-frame respect Von Neumann - Morgenstern axioms, and can therefore be represented by a utility function $\Pi_i : X \mapsto \mathbb{R}$ such that $x \succeq_i x'$ if and only if $\Pi_i(x) \geq \Pi_i(x')$. The preferences of $i$ under the I-frame should be interpreted in Savage's sense: saying that 'the person prefers $f$ to $g$' (with $f$ and $g$ two acts) means 'if he were required to decide between $f$ and $g$, no other acts being available, he would decide on $f$' (Savage, 1954, p.17). Preferences are therefore defined here as expressing choice rankings (Hausman, 2012, p.2), and should not be interpreted in utilitarian terms as expressing welfare. They therefore already integrate any empathetic concerns for others, such as a concern for fairness in the distribution of monetary outcomes. Those preferences represent the perception of the player with the I-frame (i.e. when she represents the situation of choice as a game *against*, rather than *with*, the other players): it is therefore those preferences that determine her choice if she conceives herself as an individual and not as part of a team. Furthermore, those preferences will constitute the primitive of the preferences of the individual through a we-frame. For clarity, we will refer to $\succeq_i$ as the 'I-preferences' of $i$, by opposition to the 'we-preferences' of a coalition $S_i \neq i$. We will refer to the level of utility $\Pi_i(x)$ as the *payoff* of player $i$.
Within this framework, a 'prisoner's dilemma' is a game in which the players both I-prefer the cooperative outcome to mutual defection, but also I-prefer to unilaterally defect when the other cooperates. Therefore, if P1 were able to directly choose the whole strategy profile, he will choose D for herself, and C for P2. What

makes a prisoner's dilemma a 'dilemma' is that, when both players see the game through the I-frame, they achieve *in fine* a least preferred outcome (in terms of their I-preferences) than if they had seen the game through a we-frame (leading for instance to full cooperation, which is preferred in terms of the I-preferences). Individual rational choice is here indirectly self-defeating in Parfit (1984)'s sense: we will show in section 6.5.1 that this actually provides a sufficient condition to ensure that rational players may decide to team reason — in an instrumental perspective, as a means to satisfy their I-preferences.

Borrowing Bacharach (1999)'s notion of *unreliable team interaction* (UTI), we define as a UTI a structure $\Gamma = \langle N, \{X_i\}_{i \in N}, \mathcal{S}, \mu, \Pi \rangle$ in which:

- each player $i \in N$ is characterised by a frame $S_i \in \mathcal{S}_i$ and a utility function $\Pi_i$ representing her I-preferences,
- $i$ chooses her strategy as if she were playing cooperatively with the other players in $S_i$,
- each player has a belief $\mu_i : \mathcal{S}_i \mapsto \Delta(\mathcal{S}_{-i})$ about the frame of the other players, which is derived from the common prior $\theta$ on the distribution of frames $\mathbf{S}$

The two main differences with Bacharach's notion of 'unreliable team interaction' are that (i) the players do not receive any information about the frames of the other players (we are therefore studying what Bacharach calls a 'blind UTI'), and (ii) we do not put restrictions on the team structure (the first element of Bacharach's definition of a UTI explicitly defines the set of possible teams with which the players could identify). For convenience, we assume that $\Pi_i : X \mapsto \mathbb{R}$ is $C^3$ and strictly concave in $x_i$. We introduce the following notations:

- $x_{i,S} \in X_i$ denotes the strategy of player $i$ when her frame is $S \in \mathcal{S}_i$;
- $x_i = \{x_{i,S}\}_{S \in \mathcal{S}_i} \in \{X_i\}^{2^{\mathcal{S}_i}}$ denotes the set of possible strategies of player $i$ depending on her frame;
- $x_S = \{x_i\}_{i \in S} \in \{\{X_i\}^{2^{\mathcal{S}_i}}\}_{i \in S}$ denotes the set of strategies of all the players in $S$, for all their possible frames, with $x_N = x \in X$;
- $x_{S,S} = \{x_{j,S} \in X_j \mid S_j = S\} \in X_S$ denote the strategy profile of the players $j \in S$ when their type is $S$;

- $x_{-(S,S)}$ denote the complementary profile of $x_{S,S}$, i.e. the strategies of the players $j \in N$ whose type is not $S$ (this includes the profiles $x_j$ of the non members of $S$, but also the strategies of the players $j \in S$ when their type is not $S$).

For a game in normal form $G = \langle N, X, \Pi \rangle$, $G \backslash \bar{x}_S$ denotes the game $\langle -S, X_{-S}, \bar{\Pi} \rangle$ with $\bar{\Pi}_i(x_{-S}) = \Pi_i(x_{-S}; \bar{x}_S)$, i.e. the game $G$ when the strategy of the members of the coalition $S$ is fixed to $\bar{x}_S$. Similarly, for a given UTI $\Gamma = \langle N, X, \mathcal{S}, \mu, \Pi \rangle$, $\Gamma \setminus \bar{x}_{S,S}$ denotes a UTI in which the strategy of the members of the coalition $S$ is fixed to $\bar{x}_{S,S}$.
For each UTI $\Gamma = \langle N, \{X_i\}_{i \in N}, \mathcal{S}, \mu, \Pi \rangle$, we define an associated Bayesian game $BG(\Gamma) = \langle N, \{X_i\}_{i \in N}, \mathcal{S}, \mu, U \rangle$, with $U_i : \mathcal{S} \times X \mapsto \mathbb{R}$ the utility function of player $i$ such that $U_i(x|S)$ in the Bayesian game is identical to $U_S(x)$ in the UTI, $\forall i \in S$. $BG(\Gamma)$ is therefore a game in which the types correspond to the frames of the UTI, and the individual preferences induced by a type $S$ in the Bayesian game corresponds to the we-preferences of coalition $S$ in the UTI. The crucial difference between those two games is that the players in a UTI are not necessarily expected utility maximisers. Although their frame determines their preferences, it also determines *how* they should satisfy those preferences: we will establish in the next section the precise connection between the equilibrium of $BG(\Gamma)$ and of $\Gamma$. The difference is that collective intentionality — induced by the identification to the coalition $S$ — implies that the players $i, j \in S$ are 'intending as a body' (Gilbert, 1997, 73) to satisfy their collective preferences. There is here a transfer of agency from the individual to the collective level: unlike Bayesian players who intend to satisfy collective preferences, team reasoners follow the instruction of the 'group', i.e. of a supra-individual entity who will compute the optimal strategy to satisfy the we-preferences of the team. We will refer to this abstract entity as the *supervisor* of coalition $S$.

A team reasoner $i$ therefore identifies with a coalition $S \in \mathcal{S}_i$, without being sure that the other players $j \in S$ actually identifies with the same coalition. The coalition $S$ has her own collective preferences (that we will describe in section 6.4), who are represented by a utility function $U_S : X \mapsto \mathbb{R}$. The satisfaction of those we-preferences are ensured by the supervisor of the coalition: she orders each player $j \in S$ to follow a specific strategy, such that the resulting equilibrium maximises $U_S$.

### 6.3.3 UTI and Bayesian equilibria

We now define a solution concept for a UTI, and study the relation between this concept and the equilibrium of the Bayesian game $BG(\Gamma)$. Suppose that the collective preferences $U_S$ have been defined, and that they are common knowledge among the players: each player, whatever her type is, therefore knows the objective of any possible coalition in the UTI. So as to express the idea that a team reasoner can play her part of a joint action, we introduce the relation $\sqsubset$ such that $x_{i,S} \sqsubset x_{S,S}$ should be read as 'the strategy of player $i$ when her frame is $S$ corresponds to the strategy attributed to $i$ in the profile $x_{S,S}$'.

**Definition 5.** *Let $\Gamma = \langle N, X, \mathcal{S}, \mu, \Pi \rangle$ be a UTI. A strategy profile $\bar{x} \in X$ is a UTI-equilibrium of $\Gamma$ if and only if, $\forall i \in N$:*

$$\bar{x}_{i,i} \in \arg\max_{x_{i,i} \in X_i} E\Pi_i(x_{i,i}; \bar{x}_{-(i;i)} | S_i = i), \tag{6.2}$$

$$\textit{with} \qquad E\Pi_i(x_{i,i}; \bar{x}_{-(i;i)} | S_i = i) = \sum_{S_{-i} \in \mathcal{S}_{-i}} \theta(S_{-i} | S_i = i) \Pi_i(x_{i,i}; \bar{x}_{-(i;i)}(S_{-i})), \tag{6.3}$$

$$\bar{x}_{i,S} \sqsubset \bar{x}_{S,S} \in \arg\max_{x_{S,S} \in X_S} U_S(x_{S,S}; \bar{x}_{-(S;S)}) \tag{6.4}$$

$\bar{x}$ is therefore a UTI-equilibrium if and only if I-reasoners maximise their expected utility, given the strategy of the other players, and the team reasoners $i \in S$ follow the instruction $\bar{x}_{S,S}$ that has been implemented by the supervisor of $S$ so as to maximise $U_S$ (which will be defined from the expected utility of the team members). We denote by $E[\Gamma]$ the set of UTI-equilibria of $\Gamma$.

As noted above, a central difference between a UTI and a Bayesian game is the collective rationality of teams: teams are in particular able to collectively deviate to a better equilibrium for the team, while Bayesian players with other-regarding preferences may be trapped in coordination puzzles. Consider for instance the following Hi-Lo game:

| | A | B |
|---|---|---|
| A | (R;R) | (0;0) |
| B | (0;0) | (P;P) |

with $R > P$. Denote by $\theta$ the probability of each player to be a team reasoner,

i.e. to identify with the grand coalition $N$. The expected utility of each player according to her type is therefore (with $x_{i,S} \in \{A, B\}$ the strategy of player $i$ when her frame is $S$):

$$\begin{cases} E\Pi_i(x|S_i = i) = \theta\Pi_i(x_{i,i}; x_{j,N}) + (1-\theta)\Pi_i(x_{i,i}; x_{j,j}), \\ E\Pi_i(x|S_i = N) = \theta\Pi_i(x_{i,N}; x_{j,N}) + (1-\theta)\Pi_i(x_{i,N}; x_{j,j}). \end{cases} \tag{6.5}$$

Suppose that the preferences of the grand coalition are as follows:

$$U_N(x) = E\Pi_1(x|S_1 = N) + E\Pi_2(x|S_2 = N). \tag{6.6}$$

The objective of the grand coalition is therefore to maximise the sum of payoffs of the players, but only when their type is $N$ (team reasoners are not interested in the payoff of I-reasoners). Consider the Bayesian game $BG(\Gamma) = \langle N, X, \mathcal{S}, \mu, U \rangle$ such that, $\forall i \in N$:

$$U_i(x|i) = \Pi_i(x), \tag{6.7}$$

$$U_i(x|N) = U_N(x). \tag{6.8}$$

The players share the same set of types and beliefs in $BG(\Gamma)$ and $\Gamma$, and the utility function of player $i$ when her type is $N$ corresponds to the collective preferences of the grand coalition in the UTI. We can straightforwardly show that, if $\bar{x}$ is a UTI-equilibrium of $\Gamma$, then it is also a Bayesian Nash equilibrium of $BG(\Gamma)$ (the set of UTI-equilibria of $\Gamma$ is therefore a subset of the set of Bayesian Nash equilibria of $BG(\Gamma)$). The converse is however not necessarily true: we can indeed notice that $\{x_{1,1} = B; x_{1,N} = B; x_{2,2} = B; x_{2,N} = B\}$, i.e. the strategy profile such that both players play $B$, is always an equilibrium of $BG(\Gamma)$, while this is not necessarily the case for $\Gamma$. We can indeed verify that, when we have $x_{i,i} = L \ \forall i \in N$ (I-reasoners play L), then the best reply of the coalition depends on the level of $\theta$:

$$\begin{cases} \theta > \dfrac{P}{R} & \Rightarrow \quad \bar{x}_{i,N} = A, \\ \theta < \dfrac{P}{R} & \Rightarrow \quad \bar{x}_{i,N} = B. \end{cases} \tag{6.9}$$

When I-reasoners play $B$, a team reasoner may play $A$ if the probability that the other is also a team reasoner is sufficiently high (here the threshold is $P/R$): in this situation $\{A; A\}$ is the only equilibrium, since the best reply of the I-reasoners would be to play $A$. The only situation in which $\{B; B\}$ is still an equilibrium is

when the probability that the other player also team reasons is too low. Bacharach (1999) suggested a similar result, and showed that the set of equilibria for a UTI was a subset of the set of Bayesian Nash equilibria of a Bayesian game in which types define the collective preferences of the teams of the UTI. He however did not define more precisely the adequate refinement notion: we suggest here the notion of *team-proofness* to characterise the set of relevant Bayesian equilibria.

**Definition 6.** *Let $BG = \langle N, X, \mathcal{S}, \mu, U\rangle$ be a Bayesian game. A strategy profile $\bar{x} = \{\bar{x}_{i,S_i}\}_{i\in N, S_i\in\mathcal{S}_i}$ is a team-proof Bayesian Nash equilibrium if and only if, $\forall S \in 2^N$:*

$$\forall T \subseteq S,\ \forall x_{T,S} \in X_T,\ \exists i \in T \text{ s.t. } U_i(\bar{x}|S) > U_i(x_{T,S}; \bar{x}_{-(T,S)}|S). \tag{6.10}$$

A team-proof Bayesian Nash equilibrium is therefore an equilibrium immune to collective deviations, knowing that only players of the same type can collectively deviate (it is therefore a weaker form than the strong Nash equilibrium). Two players $i$ and $j$ can indeed try to deviate together if and only if they share the same type, i.e. they identify with the same coalition $S$. We have the following result:

**Proposition 4.** *Let $\Gamma = \langle N, X, \mathcal{S}, \mu, \Pi\rangle$ be a UTI, and $BG(\Gamma) = \langle N, X, \mathcal{S}, \mu, U\rangle$ the associated Bayesian game. A strategy profile $x \in X$ is a UTI-equilibrium of $\Gamma$ if and only if it is a team-proof Bayesian Nash equilibrium of $BG(\Gamma)$.*

Proposition 4 characterises the refinement notion of Bayesian Nash equilibrium embodying a notion of collective agency: in addition of satisfying their collective preferences, the members of the same team are able to coordinate on the collectively efficient strategy profile.

## 6.4 What does 'we' want?

We now determine the we-preferences of team reasoners. We firstly highlight that a UTI can be studied as a game of complete information between coalitions, and discuss the form of the collective preferences of a coalition $S \subseteq N$ (section 6.4.1). We then characterise how team reasoners should aggregate the utilities of the members

of the coalition (subsection 6.4.2), before discussing the possibility of introducing the utilities of outsiders in the collective preferences of the coalition (subsection 6.4.3). Our analysis highlights that team reasoners are likely to be more aggressive with players outside their coalition in submodular games, whereas they will be more cooperative in supermodular games.

### 6.4.1 UTI as a game between selves

An unreliable team interaction is a game in which each player $i \in N$ identifies with a coalition $S_i$, but without necessarily knowing the frame of the others (we assume nevertheless that the types of the other players $\theta_{-i}$ are known, i.e. $i$ knows with which probability a player $j$ will identify with a coalition $S_j$). Consider a player $i$ whose frame is $S \neq i$. $i$ intends to satisfy the collective preferences of her coalition, but is not sure that the other players $j \in S$ actually identifies with $S$ (they might either be I-reasoners, or also identify with another coalition $S'$, that may partially overlap $S$, i.e. $S \cap S' \neq \emptyset$). Despite the fact that team membership is not certain, we suggest that the mere belief that the other players may be team reasoners is sufficient to ensure that a team reasoner will try to play cooperatively with the other members of her team (as in Bacharach (1999)). $i$ will however consider that there is a non null probability that $j$ does not play her part of the joint action that would be decided by the supervisor of $S$.
From $i$'s perspective, $j$ can be seen as a collection of multiples selves $(j, S_j)$, and the supervisor of $S$ must only care about the self $(j, S)$, i.e. the self of $j$ who actually identifies with $S$. It is then possible to understand a UTI as a game of complete information between the multiple selves of the different players. Formally, we will associate to any UTI $\Gamma = \langle N, X, \mathcal{S}, \mu, \Pi \rangle$ between $n$ players a game of complete information $G(\Gamma) = \langle M, X, U \rangle$ between $n2^{n-1}$ players such that:

- there exists a bijection $B : N \times \mathcal{S} \mapsto M$ such that $\forall (i, S_i) \in N \times \mathcal{S}_i$, $\exists!\ j \in M$ such that $B(i, S_i) = j$;
- $X_{i \in N} = X_{B(i,S_i) \in M}$;
- $U_{j \in M} = \sum_{S_{-i} \in \mathcal{S}_{-i}} \theta(S_{-i} | S_i = S) U_S(x_{i,S}; x_{-(i;S)}(S_{-i}))$, with $(i, S) = B^{-1}(j)$

For each combination $(i, S_i)$ in $\Gamma$, we define a unique player $j \in M$. The set of strategy of $j$ in $G(\Gamma)$ is then the same than the set of $i$ in $\Gamma$ (since $j \in M$ represents one of the selves of player $i \in N$), and the utility function of $j \in M$ is the expected utility function of $i \in N$ in $\Gamma$ when her frame is $S$. $G(\Gamma)$ is therefore

a game of complete information in which team of selves are cooperating: when the distribution of frames is realised, then only one self per player effectively plays the game — without necessarily knowing the frames of the other players. We have the following characterisation of the UTI-equilibrium of $\Gamma$:

**Proposition 5.** *Let $\Gamma = \langle N, X, \mathcal{S}, \mu, \Pi \rangle$ be a UTI and $G(\Gamma) = \langle M, X, U \rangle$ its associated game of complete information. A strategy profile $\bar{x} \in X$ is a UTI-equilibrium of $\Gamma$ if and only if, $\forall S \in 2^N$, $\bar{x}_{S,S}$ is a strong Nash equilibrium of $G(\Gamma) \setminus \bar{x}_{-(S;S)}$.*

This result directly derives from proposition 4, and means that we can define a UTI-equilibrium by directly studying the equilibria of $G(\Gamma)$: the refinement notion developed here for selecting Nash equilibria simply states that — for a given strategy profile of all the other teams $\bar{x}_{-(S,S)}$, the players in $S$ manage to coordinate on the Pareto dominant strategy profile. Note that, since all the players of $S$ have the same preferences, then we are sure that there exists at least one strong Nash equilibrium for $G(\Gamma) \setminus \bar{x}_{-(S;S)}$ (the Pareto dominant profile).

From a purely strategic perspective, a UTI can therefore be analysed as a game of complete information between coalitions (and therefore the coalition structure is commonly known). The payoff of each self $(i, S_i)$ then depends on the types of all the other players:

$$E\Pi_i(x|S_i = S) = \sum_{S_{-i} \in \mathcal{S}_{-i}} \theta(S_{-i}|S_i = S)\Pi_i(x_{i,S}; x_{-(i;S)}(S_{-i})). \tag{6.11}$$

For simplicity, we assume that the collective preferences of a team $S \subset N$ can be expressed as a weighted sum of individual payoffs, i.e. there exists real parameters $\sigma_{S,j,S_j}$ such that:

$$U_S(x|\sigma_S) = \sum_{(j;S_j) \in N \times \mathcal{S}_j} \sigma_{S,j,S_j} E\Pi_j(x|S_j), \tag{6.12}$$

with $\sigma_{S,j,S_j}$ the weight coalition $S$ attributes to player $j$ when her frame is $S_j$. The sign of $\sigma_{S,j,S_j}$ therefore indicates whether the coalition $S$ tries to cooperate or not with player $j$ when her frame is $S_j$. The object of this section is to determine the value of the weights $\sigma_{S,j,S_j}$.
We will firstly deal with the aggregation of individual preferences within the same

coalition, before treating the possibility of presenting cooperative or aggressive preferences towards outsiders of the coalition. A core element of our model is that we assume that those weights are chosen *strategically*: rather than imposing the collective preferences, we consider that the team reasoners will seize the opportunity to choose their collective preferences so as to maximise *in fine* their own payoff.

Similarly to the model developed in chapter 5, for each UTI $\Gamma$, we define a two-stage game $\Gamma^*$ as follows:

- the second stage $G(\Gamma|\sigma)$ is the game in normal form $G(\Gamma)$ associated to $\Gamma$ such that the weights of any utility function $U_S(.|\sigma_S)$ are given by $\sigma \in \mathbb{R}^{(2^n-n-1)\times n2^{(n-1)}}$;
- the first stage game $G_0(\Gamma) = \langle M, \mathbb{R}^{(2^n-n-1)\times n2^{n-1}}, \Pi\rangle$ is a game in normal form in which each player of the $(2^n - n - 1)$ possible coalitions (the number of sets of $N$ of at least two elements) chooses $n2^{n-1}$ weights to define their collective preferences.

Team reasoners therefore choose their own collective preferences in the first stage game $G_0(\Gamma)$ (I-reasoners are simply maximising their expected utility, since they do not have the opportunity to choose collective preferences), and then play the UTI with those specific collective preferences. Since team reasoners are actuated by a collective objective, we cannot simply assume that team reasoners choose the weights $\sigma$ in the first stage game so as to maximise their expected payoff (two team reasoners of the same team would then be likely to choose different weights, and therefore would not have the same collective preferences). We will discuss this point in more details in the section 6.4.2.
Finally, so as to alleviate the presentation of our results — and for similar reasons than the ones exposed in chapter 5 — we assume that, $\forall\sigma \in \mathbb{R}^{(2^n-n-1)\times n2^{(n-1)}}$, $G(\Gamma|\sigma)$ has a unique Nash equilibrium $\bar{x}(\sigma)$, i.e. $\exists!\bar{x}(\sigma) \in X$ such that, $\forall j \in M$:

$$\frac{\partial U_j}{\partial x_j}(\bar{x}(\sigma)|\sigma) = 0, \tag{6.13}$$

$$\frac{\partial^2 U_j}{\partial x_j^2}(\bar{x}(\sigma)|\sigma) < 0. \tag{6.14}$$

We could otherwise ensure the existence of an equilibrium by allowing for mixed strategies (the collective utility functions are indeed continuous by construction), and many Nash equilibria would not actually satisfy the conditions of proposition 5.

### 6.4.2 Within-group preferences

We assumed that team reasoners are able to strategically choose their own collective preferences. We therefore defined a two stage game $\Gamma^*$ in which team reasoners choose the weights $\sigma$. The last issue to be dealt with is to define the criterion of selection of those weights, i.e. what objective a team reasoner should pursue when choosing the form of her collective preferences, knowing that the other members of her team are pursuing the same objective. This question consists in knowing how the team reasoners should aggregate their preferences within their own team[9].

Bacharach suggests the principle of 'Paretianness' for describing the collective preferences $U_S$ of the team: if all the players I-prefer $x$ to $x'$, then $U_S(x) \geq U_S(x')$. Since many functions respect this condition, it is not certain that all the players will perceive the same collective objective, in particular when their interest are in conflict. Bacharach (2006, 88) suggests then that 'in circumstances in which nothing is perceived by individual members about other individual members beyond the facts recorded in a bare game representation, principles of fairness such as those of Nash's axiomatic bargaining theory will be embedded in $[U_S]$'. Bacharach's approach for the determination of we-preferences is axiomatic: he suggests stating *a priori* certain desirable properties of fairness and efficiency, and then identifying whether collective preferences respecting those properties may exist. Defining collective preferences as the result of a bargaining process seems quite intuitive, since it means that the players would all agree on those collective preferences. Team reasoners indeed ask themselves which strategy profile they should choose collectively so as to satisfy the interest of each team reasoner: they are therefore trying to select a profile that Pareto dominates their disagreement payoff (the payoff they would achieve if they were I-reasoning) with which everyone would agree.
In this chapter, we will assume that team reasoners use the Nash solution as a solution concept, and that it is common knowledge among all players[10]. We define

[9]Note that we are studying the game of complete information associated to $\Gamma$, so the teams are not unreliable any more: each team reasoner knows that the other members of the teams (the selves $(j;S)$) are also team reasoning with $S$.

[10]A difficulty of the axiomatic approach is that there are very few unquestionable axioms such as Paretianness, and it will probably be necessary to state more controversial axioms to be able to isolate a restricted number of solutions (such as the independence of irrelevant alternatives (Nash, 1950)). We will however not discuss further whether there is a more relevant set of axioms than another one, and simply assume that it is common knowledge that the team reasoners refer to Nash (1950)'s axioms. The main contribution of our work is indeed not about the aggregation of individual preferences within a team, but about the formation of collective preferences with respect to other players. We will indeed show in section 6.5 that it is the possibility to choose aggressive

the disagreement payoff of coalition $S$ as the payoff its members would get if they were not team reasoning, i.e. if they were simply maximising their individual utility. It is therefore a UTI-equilibrium $\bar{x}$ in which $\bar{x}_{S,S}$ does not maximise the collective preferences $U_S$ of coalition $S$, but each $\bar{x}_{i,S}$ maximises $E\Pi_{i,S}$ (knowing that the other teams $S'$ are maximising their collective preferences $U_{S'}$ and the I-reasoners are maximising their expected utility $E\Pi_{j,j}$). For convenience, we will assume that the disagreement payoff is common knowledge: although several equilibria may exist, all the team reasoners agree on what would be the non-cooperative outcome.
The team reasoners must therefore choose collective preferences so as to maximise *in fine*:

$$\Pi_S(x) = \prod_{i \in S} \left[ E\Pi_{i,S}(x_{S,S}; x_{-(S,S)}) - E\Pi_{i,S}(\bar{x}) \right]. \tag{6.15}$$

The team reasoners therefore aggregate their preferences such that the resulting equilibrium maximises the joint product of the gains that each individual gets from being a team reasoner. We can then notice that Harsanyi (1955)'s theorem implies that there exists real parameters $\sigma_{S,i,S}$ such that maximising $\Pi_S$ is equivalent to maximising:

$$\bar{U}_S(x) = \sum_{i \in S} \sigma_{S,i,S} E\Pi_i(x_{S,S}; x_{-(S,S)}). \tag{6.16}$$

For instance, if the team is symmetric (the players have the same set of strategies, and we can permute the indices of the expected utility functions), then those weights will be equal — due to the axiom of symmetry characterising the Nash solution. On the contrary, non-symmetric games will generally lead to non-egalitarian weights. We can therefore represent the within-group preferences of $S$ as a weighted sum of the expected utility of each member of the team.

### 6.4.3 Between-group preferences

We have determined in the previous section that team reasoners should choose their collective preferences so as to maximise *in fine* $\bar{U}_S$. We show in this section that it is generally in the interest of the team reasoners to choose weights such that $U_S \neq \bar{U}_S$. Similarly to our analysis of chapter 5, we will see that the coalition $S$ chooses its

or cooperative preferences toward outsiders that will justify the emergence of team reasoning in non-cooperative games.

collective preferences $U_S$ so as to reach a Stackelberg equilibrium (given its true payoff $\bar{U}_S$). It may therefore be in the interest of $S$ to be aggressive or cooperative with other players. The only differences with the model we developed in chapter 5 are that (1) I-reasoners do not have the opportunity to choose collective preferences, and are therefore committed to the maximisation of their individual payoff, and (2) preferences are chosen at a collective level. Similarly to chapter 5, we have more parameters to determine than first order conditions ($(n2^{n-1}-|S|)$ parameters for $|S|$ conditions), and several specifications of the optimal weights are possible. However, and unlike proposition 2, it will not be possible to determine a relatively 'clear' and general presentation of the weights. We therefore state the following proposition by focusing on aggregative games, so as to be able to offer a presentable solution (the general solution is detailed in appendix):

**Proposition 6.** *Let $\Gamma$ be an aggregative UTI, i.e. an UTI such that $\Pi_i(x) = \Pi_i(x_i; \sum_{j\neq i} x_j)$. The collective utility function $U_S$ of a coalition $S$ is:*

$$U_S = \bar{U}_S(x) + \sum_{j\in -S} \bar{\sigma}_{S,j,S_j} \bar{U}_{S_j}(x), \tag{6.17}$$

$$\text{with} \qquad \bar{\sigma}_{S,j,S_j} = \frac{\frac{\partial \bar{U}_S}{\partial x_{j,S_j}}}{\frac{\partial \bar{U}_{j,S_j}}{\partial x_i}} \frac{C_{ij}^{J(S)}}{C_{ii}^{J(S)}}, \tag{6.18}$$

*with $J(S)$ a $(m-|S|+1)\times(m-|S|+1)$ matrix identical to $J(G(\Gamma))$ in which all the rows and columns $j \in S \setminus i$ have been deleted, and:*

$$\bar{U}_S(x) = \sum_{i\in S} \sigma_{S,i,S} E\Pi_i(x_{S,S}; x_{-(S,S)}), \tag{6.19}$$

*with the weights defined by (6.16).*

Proposition 6 gives the expression of the optimal weights a coalition $S$ should give to the other players so as to maximise the joint product of the individual profits from team reasoning. The question that follows is then to determine under which conditions the weights $\bar{\sigma}_{S,j,S_j}$ are positive or negative, i.e. under which conditions it is in the interest of the coalition $S$ to be more cooperative or more aggressive with outsiders.

Note that proposition 6 is extremely similar to proposition 2, since in both cases

an agent (a single player or a coalition) is choosing her optimal weights given her objective function (her material payoff, or the joint product of the benefice of team reasoning). We have therefore here the same result concerning the sign of the optimal weights: since the coalitions try to reach a Stackelberg equilibrium, they will choose cooperative preferences in supermodular games (since being more cooperative will generate a positive best reply from the others), while they will choose aggressive preferences in submodular games. This means that, in addition of aggregating the preferences within their team, team reasoners are likely to choose collectively whether they are cooperative or competitive with outsiders: in environments characterised by strategic substitutes, group of individuals will tend to enter in competition with other groups so as to maximise the difference in payoff before their groups, whereas environments characterised by strategic complementarities will generate more cooperative relations between the different groups. Our model of team reasoning may therefore give insights about the formation of coalition and in particular of the emergence of cooperation and competition between coalitions.

## 6.5 Why should we team reason?

We have characterised in the previous section the optimal behaviour of team reasoners for a given distribution of types. We now suggest endogenising the types of the players, by studying the conditions under which the players are actually better off when they team reason. We argue that, under the assumption of common knowledge of rationality, rational players are able to directly choose their own frames (section 6.5.1), and show that it is generally in the interest of the players to choose to team reason (section 6.5.2).

### 6.5.1 Choosing one's frame

As described in the previous section, team reasoning implies that the players must choose collective preferences so as to satisfy *in fine* their own I-preferences. It therefore offers them an opportunity to make a strategic commitment: if players were able to choose their frames, it could then be in their interest to identify with a group $S$, such that the satisfaction of their we-preferences leads to an equilibrium in which their I-preferences are better satisfied. Since our purpose is to offer a normative theory of rational choice, we must investigate whether rational players ought to choose their own frames. The players would then face a three-stages game: (i) they firstly choose their frame $S$, (ii) they then choose the optimal preferences

for $S$ (the weights $\sigma$), and (iii) they collectively satisfy the preferences they have chosen in (ii).

For a game $G = \langle N, X, \Pi \rangle$ in normal form, we can define a first stage game $G_0 = \langle N, \mathcal{S}, V \rangle$ such that:

$$V(\mathbf{S}) = \Pi(\bar{x}(\mathbf{S})), \tag{6.20}$$

with $\bar{x}(\mathbf{S})$ the UTI equilibrium of $\Gamma = \langle N, X, \mathcal{S}, \mu, \Pi \rangle$, with $\theta(\mathbf{S}) = 1$, i.e. the distribution of frames $\mathbf{S}$ is common knowledge. $G_0$ is therefore a game in which each player $i \in N$ chooses her frame $S_i$ so as to maximise her utility $\Pi_i$, knowing that choosing $S_i$ when the others choose the frames $S_{-i}$ implies that they must play a UTI-equilibrium. As an illustration, consider a prisoner's dilemma:

| | C | D |
|---|---|---|
| C | (R;R) | (S;T) |
| D | (T;S) | **(P;P)** |

with $T > R > P > S$, and $2R > T + S$. The unique Nash equilibrium of the game is $\{D; D\}$, whose payoff is $(P; P)$. Suppose now that P1 and P2 are team reasoners, i.e. $S_i = N$. In this situation, we can easily check that the Nash solution to the bargaining problem faced by P1 and P2 is simply $(R; R)$, i.e. both players should cooperate. Suppose now that P1 is a team reasoner, while P2 is not (and this is common knowledge). As a utility maximiser, P2 therefore defects. As a team reasoner, P1 wants to choose collective preferences such that all the team reasoners (i.e. only herself in this situation, since the probability that P2 team reasons is null) agree with the resulting equilibrium. P1 therefore also defects.
Suppose now that both players are able to directly choose their frames. The first stage game can be represented by the following payoff matrix:

| | $S_2 = \{2\}$ | $S_2 = N$ |
|---|---|---|
| $S_1 = \{1\}$ | **(P;P)** | (P;P) |
| $S_1 = N$ | (P;P) | **(R;R)** |

The first stage game of the choice of one's frames in a prisoner's dilemma has therefore two Nash equilibria: being both I-reasoners, or being both team reasoners. We can however notice that the frame $S_i = N$ weakly dominates the frame $S_i = i$, $\forall i \in N$. In a prisoner's dilemma, both players has therefore a strict incentive in

becoming team reasoners if they believe that the other is likely to be a team reasoner.

We can now provide an equilibrium notion integrating the possibility to choose one's own frame in the first stage game $G_0$:

**Definition 7.** *Let $G = \langle N, X, \Pi \rangle$ be a game in normal form. A strategy profile $(\bar{x}; \bar{\boldsymbol{S}}) \in X \times \mathcal{S}$ is a subgame perfect UTI-equilibrium if and only if:*

- *$\bar{x}$ is a UTI-equilibrium of $\Gamma = \langle N, X, \mathcal{S}, \bar{\mu}, \Pi \rangle$, with $\bar{\mu}$ the beliefs induced by the distribution of frames $\bar{\boldsymbol{S}}$;*
- *$\bar{\boldsymbol{S}}$ is a Nash equilibrium of $G_0 = \langle N, \mathcal{S}, V \rangle$, with $V_i(\boldsymbol{S}) = \Pi_i(\bar{x}(\boldsymbol{S}))$.*

A subgame perfect UTI-equilibrium is therefore a combination of a UTI equilibrium $\bar{x}$ and a distribution of frames $\bar{\mathbf{S}}$ such that the frames have been strategically chosen. There exist for instance two subgame perfect UTI-equilibria in a PD, $(\{C; C\}; \{N; N\})$ and $(\{D; D\}; \{i; i\})$: the players can either be both team reasoners and cooperate, or be both I-reasoners and defect.

We can notice that a subgame perfect UTI-equilibrium is an equilibrium notion for games in normal form, and not for a UTI (the UTIs are defined so as to determine the optimal choice of the players when they decide to team reason). Since our purpose is to offer a theory explaining how rational players ought to choose, we can wonder under which conditions the choice of rational players constitute a subgame perfect UTI-equilibrium rather than a Nash equilibrium. We argue here that the assumption of common knowledge of rationality constitutes a sufficient reason for rational players to team reason, since it provides them a reason to make a strategic commitment.
As a guiding illustration, consider the symmetric Cournot game we mentioned earlier with two players. As a rational player, P1 knows that Bayesian rationality may be indirectly self-defeating, and therefore that she may benefit from a strategic commitment if P2 knows that P1 is making a commitment (P1 would indeed benefit from an aggressive behaviour, since she may reach *in fine* her Stackelberg equilibrium). The central point of our argument is that assuming CKR is sufficient to ensure that P2 may rationally believe that P1 makes a commitment, and therefore that P1 have an incentive in effectively making a commitment. P1 and P2 indeed know

that Bayesian rationality is indirectly self-defeating, and this is common knowledge — it is indeed a mathematical property of the payoff structure. P1 therefore knows that she may benefit from presenting aggressive preferences if P2 believes it. The only reason why P2 would believe that P1 makes a strategic commitment is that P1 believes that P2 believes it (and hence makes a strategic commitment). By using an infinite-regress based argument (such as the one supporting the choice of strategy B in a Hi-Lo game), we can find a reason for $i$ to make a strategic commitment, i.e. the common belief that $i$ makes a commitment.

Since rational players know that they can benefit from strategic commitments if the others know that they make a commitment, and that they may rationally construct the beliefs that the other effectively knows that they make a strategic commitment when they make a strategic commitment, the players have the opportunity to choose in a first stage game their optimal commitment. Within our framework, this means that players are able to rationally choose their own frames, so as to satisfy *in fine* their I-preferences[11]. Rational players therefore choose their frames in the first stage game, knowing that a given distribution of frames in the first stage game will induce a UTI of complete information (i.e. the frames are perfectly observed) in the second stage. Their final choice therefore constitutes a subgame perfect UTI-equilibrium: if the structure and the Bayesian rationality of the players is common knowledge, then the players' choice $\bar{x} \in X$ is such that there exists $\bar{\mathbf{S}} \in \mathcal{S}$ such that $(\bar{x}; \bar{\mathbf{S}})$ is a subgame perfect UTI-equilibrium.

An interesting implication of this result is that, in a prisoner's dilemma, rational players are not condemned to defect: they can also choose to become team reasoners and *in fine* cooperate. The underlying logic of this result is that each player decides to become a conditional cooperator: although it is preferable for an individual to systematically defect in front of a systematic cooperator, conditional cooperators can be more successful than unilateral defectors because they achieve a higher payoff when playing with other conditional cooperators. As discussed in chapter 4, our model of team reasoning is very similar to the idea of constrained maximisation suggested by Gauthier (1986): indeed, team reasoners agree on the principles that should guide their conduct (cooperate if and only if the other cooperates), and choose those principles so as to satisfy *in fine* their own interest.

[11] We will show in chapter 7 that the possibility of team reasoning can actually be seen as the optimal form of strategic commitment, since it allows for the formation of an optimal individual commitment (as in chapter 5) when players unilaterally team reason, and also for collective optimal commitments.

### 6.5.2 Optimal frames

We now characterise the properties of the Nash equilibrium of the first stage game $G_0$. We will study in particular the conditions under which $S_i = i$, $\forall i \in N$ is a Nash equilibrium of $G_0$. Those conditions indeed characterise the set of games for which the players prefer to be I-reasoners rather than team-reasoners.
Consider a n-player game in normal form $G$. Suppose that the structure of the game is common knowledge as well as the rationality of the players: the players therefore choose their frames $S_i$ so as to reach a Nash equilibrium of $G_0$. We have the following result:

**Proposition 7.** *$\{\bar{S}_i = i\}_{i \in N}$ is a Nash equilibrium of $G_0 = \langle N, \mathcal{S}, V \rangle$ if and only if, $\forall i \in N$:*

$$\Psi_i(\bar{x}_i) \geq \Psi_i(x_i), \qquad \forall x_i \in X_i. \tag{6.21}$$

This result simply states that, as soon as a player can benefit from becoming a Stackelberg leader, then it is in her interest to become a team reasoner. The basic idea of the proof is that, knowing that all the other players $j \neq i$ are I-reasoners, if $i$ decides to identify with $S_i \neq i$, she plays a bargaining game with the other team reasoners. Since she is the only team reasoner, she does not need to care about the utility of the other members of $S_i$, and simply chooses the we-preferences that maximise her utility. This is the precise definition of a Stackelberg equilibrium for $i$.
It is therefore generally in the interest of player $i$ to become a team reasoner. Notice now that, if $i$ chooses to deviate to the frame $S_i = \{i, j\}$, then player $j$ is likely to be better off by becoming a team reasoner too, since she will now negotiate the preferences of the coalition with $i$: teams are therefore likely to emerge in the game $G_0$. The existence of $n$ players however makes the characterisation of the equilibrium quite delicate, since many coalitions are likely to emerge simultaneously. It would however be possible to study the possible coalition structures at equilibrium by appealing to notions of coalitional stability (e.g. d'Aspremont et al. (1983) on cartels, Barrett (1999) on self-enforcing agreements, Bloch et al. (2006) on contests games). We suggest therefore focusing on two-player games in order to derive more complete results:

**Proposition 8.** *Let $G$ be a game in normal form with $n = 2$. Then:*

*(i)* $\{\bar{S}_i = N\}_{i \in N}$ *is always a Nash equilibrium of* $G_0$,

*(ii) there is no subgame perfect equilibrium* $(x; \boldsymbol{S})$ *such that* $x \neq \bar{x}$ *if and only if (i) no player can benefit from a first mover advantage, (ii)* $\bar{x}$ *is Pareto optimal with no Pareto equivalent profile (i.e. a profile* $x$ *such that* $\Pi_i(x) = \Pi_i(\bar{x})$ $\forall i \in N$*)*

Claim (i) states that, in two-player games, being both team reasoners is always a Nash equilibrium in the first stage game. This implies that there always exists a subgame perfect equilibrium in which both players team reason in two-player games: under the assumption of CKR, rational players in a game in normal form $G$ should always be able to reach the Nash solution of $G$ (if $G$ was played cooperatively). The fundamental intuition of this result is that, since I-reasoning may be indirectly self-defeating, players intending to maximise their expected utility should decide in a first step *how* they want to choose their optimal strategy (by maximising their expected utility, or by following another rule), before implementing the optimal strategy according to this optimal rule of decision. Proposition 8 states that being a team reasoner is always an optimal rule of decision.
Claim (ii) determines the necessary and sufficient conditions under which rational players cannot reach an other outcome than Nash equilibrium. The two conditions are simply that (i) no player has an interest in becoming a Stackelberg leader, and (ii) the players do not have the incentive to deviate collectively to a Pareto superior (or equivalent) profile. This result means that, in two-player games, unless the Nash equilibrium offers to both players their Stackelberg payoff and is Pareto optimal, then players can rationalise the choice of another strategy profile by becoming a team reasoner. For instance, in the PD, although the Nash equilibrium pays to both players their Stackelberg payoff, it is not the unique equilibrium that can be implemented in a subgame perfect UTI-equilibrium — it is indeed not Pareto optimal.
Several configurations are therefore possible in two-player games:

- if the Nash equilibrium is not the Stackelberg equilibrium of one of the players, then this player has in interest in becoming a team reasoner; the second player has then an interest in becoming a team reasoner too (they can either both team reason and achieve the same Stackelberg equilibrium, or achieve an equilibrium in which the second player is better off);

- if the Nash equilibrium is the Stackelberg equilibrium of both players, then being a I-reasoner and playing a Nash equilibrium is a subgame perfect UTI-equilibrium; if the Nash equilibrium is not Pareto optimal, then there exists a second subgame perfect UTI-equilibrium, in which both players team reason and achieve an outcome that Pareto dominates the Nash equilibrium;

- if the Nash equilibrium is the Stackelberg equilibrium of both players and is Pareto optimal, then the Nash equilibrium is the only equilibrium implementable in a subgame perfect UTI-equilibrium.

## 6.6 Conclusion

In this chapter, we argued in favour of team reasoning as the foundation of an economic theory of unselfish behaviour. We firstly showed that standard game theory faces some puzzling results in cooperation and coordination games due to its very peculiar conception of strategic interactions as a simple extension of individual rational choice theory. We suggested that modelling strategic interactions requires introducing a notion of collective agency: players are not isolated expected utility maximisers, but consider each other as *agents*, able to choose what matters for *them* (their collective preferences), and not as simple parameters descriptive of the state of the world. This approach probably offers a better account of what strategic interactions actually are, since it better captures the relational nature of strategic interactions which is lacking in standard game theory. Team reasoning indeed account for the relational nature of humankind (Bruni and Sugden, 2008): social interactions, unlike within standard economic theory where the only motive of action is the satisfaction of one's own preferences, can be understood as joint intentions for mutual assistance.

We then developed a formal model of team reasoning based on Bacharach's work, and suggested a model of collective preferences formation. We argued that collective preferences are the result of a strategic choice and therefore that they can be defined from solution concepts of bargaining games. We then showed that it is generally in the interest of team reasoners to choose collective preferences that integrate the utility of outsiders. Choosing as a member of a team defines the way we interact with players we do not identify with: although I individually do not care about my neighbour, the construction of my collective preferences (and then of my own identity, as the member of a group) may lead me to be aggressive or cooperative with her. Since becoming a team reasoner gives the opportunity to the individuals to choose what matters for them, as a group, it gives them the

opportunity to make strategic commitments. We then showed that it is actually generally in the interest of the players to become team reasoners.

We would like to conclude this chapter by stressing some implications of our model for the reconciliation problem. Conventional welfare economics assumes that individuals act as if seeking to satisfy coherent preferences, and take the satisfaction of those preferences as the normative criterion. Team reasoning however allows for a transfer of agency from the individual to the collective level, and we showed that this procedure can be analysed as the formation of collective preferences. We can therefore wonder which preferences should be taken into account when producing normative assessments: the initial ones representing the individualistic assessment of the states of the world, or the final ones integrating the social dimension of the individual and her relations towards others? We can indeed notice that, from the perspective of variable frame theory, it is not clear whether the satisfaction of our I-preferences matters more than the satisfaction of our we-preferences (they are indeed both valid perceptions of the same state of the world). Furthermore, we showed in this chapter that intending to satisfy one's preferences may lead to self-defeating behaviours: it is therefore in the interest of the individuals to choose instrumental preferences so as to be able to satisfy their initial preferences.
Accepting variable frame theory questions the validity of preference satisfaction as a normative criterion, since 'individual preferences' cannot be unambiguously defined for each player. What matters for the individual is indeed determined by how she perceives the world and her relation with others (whether they are members of a common group or not). Instead of seeing the individuals as passive utility maximisers, our analysis suggests that the individuals can also choose their own preferences. This new perspective may lead to an important change in normative economics and the way economists think of public policies: instead of designing incentives for steering rational individuals into a given direction (i.e. modifying the payoff of the game to ensure that the resulting Nash equilibrium corresponds to what the social planner intends to achieve), it could be more relevant to help the individuals to engage in team reasoning (by facilitating the communication between them for instance). Within the context of the management of CPR discussed in chapter 4, a possible interpretation of Ostrom (1990)'s design principles would be that governments should ensure to the individuals the means to choose together their own collective preferences: inducing the individuals to perceive a situation with a we-frame instead of a I-frame may therefore offer an alternative solution to deal with collective action issues. Nagatsu (2015), referring to the anti-littering campaign 'Don't mess with Texas', argues for instance in favour of *social nudges*

as a way to influence how people frame the game: Nagatsu argues that, since the framing of the choice problem occurs *before* the exercise of practical reasoning, then the nudge is not manipulative (it does not use deliberatively a flaw in human decision making), and respect individual autonomy. If a nudge is likely to extend the set of frames the individual has at her disposal (she may for instance become aware that she can conceive herself as playing a game with others rather than against them), then it would be compatible with our normative criterion of individual autonomy — it would indeed contribute to increase the autonomy of the individual by making her aware of the influence her frame may have on her final choice.

Chapter 7

# The Ecological Rationality of Team Reasoning

## Contents




**Abstract:** We provide in this chapter a justification of team reasoning as an ecologically rational heuristic. We slightly amend the indirect evolutionary approach — by suggesting studying the evolution of individual heuristics rather than individual preferences — and study the evolutionary stability of payoff maximising behaviour in finite games. We define the strict $\phi$-core as a refinement of the $\gamma$-core by assuming that the deviating coalition acts as a Stackelberg leader and the remaining individuals as singletons. We show that maximising one's payoff is evolutionary stable if and only if the resulting Nash equilibrium belongs to the strict $\phi$-core. We can then highlight that, when the players use the heuristics that outperform payoff

maximisers, they behave as if they were team reasoning. This result suggests that, from an evolutionary perspective, team reasoning is ecologically rational: although individual behaviour differs from the prescription of the "constructivist rationality" of neoclassical economics, individual heuristics consistent with team reasoning are well adapted to the environment within which the players interact.

## 7.1 Introduction

A core hypothesis of economic theory is that individuals behave as if seeking to satisfy their preferences by maximising their expected utility (Friedman, 1953). The assumption of payoff maximising behaviour has generally been justified by evolutionary arguments when studying competitive markets (Alchian, 1950, Friedman, 1953) and empirical results suggest that individuals effectively converge to such behaviours in repeated market experiments (Plott, 1996). However, a large literature has shown that departures from payoff maximisation are likely to survive evolutionary pressures, such as concerns about fairness (Güth and Yaari, 1992, Huck and Oechssler, 1998), altruism (Bester and Güth, 1998), spite (Bolle, 2000) or competition and concern for relative success rather than absolute success (Kockesen et al., 2000a,b).

The aim of this chapter is to determine in a quite general setting necessary and sufficient conditions under which individuals will effectively converge to a payoff maximising behaviour (henceforth 'PMB') in the long run. Unlike the indirect evolutionary approach (Güth and Yaari, 1992, Güth, 1995), we suggest studying the evolution of individual *heuristics* (defined as a procedure that associates to any possible situation the choice to make) rather than of individual preferences. We indeed argue that this approach provides more complete predictions, and probably constitutes a more natural extension of standard evolutionary game theory. Furthermore, it provides a simple framework to catch the idea of ecological rationality suggested by Simon (1955, 1956), developed by Gigerenzer et al. (1999), Gigerenzer and Selten (2001) with the notion of "fast and frugal heuristics", and by Smith (2003, 2008), according to whom individual rationality in markets is the result of the combination of constructivist rationality (rationality in the traditional sense, that Davis (2011, p.150) defines as "conscious, deliberative, and deductive — in a word, Cartesian") with the ecological rationality of the institution of markets. We indeed develop a framework in which no notion of "true preferences" is necessary – in line with Simon's rejection of utility functions —, whose primitive

is the existence of a criterion for evolutionary selection.

We borrow from cooperative game theory the notion of $\gamma$-core (Chander and Tulkens, 1997) and refine it with the notion of $\phi$-core (Currarini and Marini, 1998) and strict $\phi$-core: the $\phi$-core is defined as the set of payoff vectors such that no coalition can unilaterally benefit from deviating from the agreement, knowing that the deviating coalition acts as a Stackelberg leader and the remaining individuals maximises their individual payoff as singletons. The strict $\phi$-core rules out payoff vectors that a coalition $M \subset N$ could have achieved on its own. Our main proposition states that PMB is evolutionary stable if and only if the resulting Nash equilibrium belongs to the strict $\phi$-core. This result means that, as soon as a coalition can benefit from a first mover advantage, then some players may benefit from not playing their best reply and from acting "irrationally". Since the set of games for which there exists a Nash equilibrium in the strict $\phi$-core is quite restrictive (mainly zero-sum games and games with a Pareto dominant outcome), the soundness of the as-if hypothesis — and in particular its evolutionary justification — can be seriously questioned. Furthermore, we will show that players following the heuristics that outperform PMB behave as if they were team reasoning: it means therefore that evolutionary pressures are likely to select team reasoners, since their mode of reasoning is more adapted to the environment within which they interact.
The rest of this chapter is organised as follows. In section 7.2, we focus on a prisoner's dilemma in order to discuss the indirect evolutionary approach and our reformulation in terms of evolution of heuristics. In section 7.3, we define our formal framework. Section 7.4 states our main theorem. Section 7.5 concludes.

## 7.2 Cooperation in a prisoner's dilemma

The indirect evolutionary approach suggests studying the evolution of individual preferences according to the material payoff they generate: unlike usual evolutionary game theory, in which it is assumed that behaviours are genetically or phenotypically determined (Hammerstein and Selten, 1994, Boyd and Richerson, 1985), we do not directly study the evolution of the actual behaviour of the individuals, but the evolution of their preferences. It is assumed that the preferences of the players are defined at each date by a specific payoff distortion: their behaviour may therefore indirectly evolve *via* the evolution of their underlying preferences. Formally, we model a "preferences game", in which individual payoffs for each combination of individual preferences correspond to the Nash equilibrium in the normal form game

defined by those preferences. There then exists an evolutionary dynamics that selects the individuals according to the payoffs generated by their types, i.e. according to the payoff distortion characterising their preferences. The indirect evolutionary approach consists in the study of the evolution of individual preferences in this (fictive) preferences game.
The different results mentioned above suggest that PMB does not probably constitute an evolutionary stable strategy of the preferences game. Heifetz et al. (2007a,b) conduct for instance an analysis of preference evolution by seeking general conditions under which departures from PMB may survive evolutionary pressures: they show that some degree of payoff distortion is generically beneficial to the players, and therefore that the population's type cannot converge to PMB. Furthermore, in the case of symmetric two-player games, they determine sufficient conditions (mainly dominance solvability of the preferences game) under which the population effectively converges to the Nash equilibrium of the preferences game. However, although PMB does not seem to be an equilibrium of the preferences game, the mere *nature* of the preference distortion is not well established: Heifetz et al. (2007b)'s disposition function for instance captures a wide range of distortions, such as altruism, spite or social status, but they are restricted to *payoff* distortions rather than *preference* distortions in general. In particular, it is not certain that payoff distortions may easily capture any possible preference distortion: the specific kind of distortion function we choose for studying the evolution of preferences may have a great impact on the possible outcomes of the evolutionary process.
As an illustration of this point, we will study the following prisoner's dilemma (PD):

| | C | D |
|---|---|---|
| C | $(R;R)$ | $(S;T)$ |
| D | $(T;S)$ | $(P;P)$ |

with $T > R > P > S$ and $2R > T + S$. We suggest studying whether cooperating is an evolutionary stable strategy, according to the trait that is subject to evolutionary pressures: strategies, heuristics or payoff distortions. The aim of this illustration is to highlight a fundamental weakness of the indirect evolutionary approach, and to argue in favour of a game-theoretical analysis of evolutionary processes in terms of heuristics rather than payoff distortions.

### 7.2.1 Cooperation in standard evolutionary game theory

Suppose that there exists a population $[0;1]$ from which two individuals are randomly drawn to play this game. In evolutionary game theory, we assume that this

population is composed of two types of individuals, cooperators and defectors: the former type is programmed to cooperate whatever happens while the latter systematically defects (each type therefore corresponds to a possible strategy). We then assume that a process of natural selection operates within the population, such that the share of a specific type increases if and only if it generates higher material payoff when randomly matched with another individual from the population. The level of payoff associated to a specific type defines the *fitness* of the individual holder of this type. Formally, the share of a type $i$ evolves as follows:

$$p_i^{t+\mathrm{d}t} = \frac{p_i^t \Pi_i^t}{\bar{\Pi}^t}, \tag{7.1}$$

with $\Pi_C^t$ (resp. $\Pi_D^t$) the expected payoff of a cooperator (defector) at the date $t \geq 0$, and $\bar{\Pi}^t$ the average expected payoff:

$$\Pi_C^t = S + p_C^t(R - S), \tag{7.2}$$
$$\Pi_D^t = P + p_C^t(T - P), \tag{7.3}$$
$$\bar{\Pi}^t = p_C^t \Pi_C^t + (1 - p_C^t)\Pi_D^t. \tag{7.4}$$

Equation (7.1) means that the share of cooperators in $t+\mathrm{d}t$ is equal to its "contribution" to the average payoff[1] in $t$. It follows that the proportion of cooperators can increase if and only if $\Pi_C^t > \bar{\Pi}^t$. We thus obtain the standard replicator dynamics equation (Maynard Smith, 1982):

$$\dot{p}_i = p_i^t \left[ \frac{\Pi_i^t - \bar{\Pi}^t}{\bar{\Pi}^t} \right]. \tag{7.5}$$

The evolution of types in this game is therefore deterministic. This dynamics can be understood in a biological sense, as the selection of the most fitted genotypes, but this kind of process is only effective in the very long term. For economic analysis, it is probably more relevant to interpret such evolution as the result of learning and education (Weibull, 1995, section 4.4): among a population of identical individuals, if one player follows a strategy that generates higher payoffs than the others, then

[1] A possible justification of this relation is to refer to the interpretation of evolutionary biology: if we consider a population of $n$ individuals, and interpret $\Pi_i^t$ as the number of offspring of the individuals of trait $i$ at date $t$, then $n\bar{\Pi}^t$ denotes the total number of offspring at date $t$. The frequency of individuals of trait $i$ in $t + \mathrm{d}t$ (i.e. among the offspring of period $t$) is therefore the ratio between the number of offspring of trait $i$ at date $t$, i.e. $np_i\Pi_i^t$, and the total number of offspring at date $t$, $n\bar{\Pi}^t$. We then obtain the relation (7.1).

the less successful players might be tempted to imitate her.
We can easily see that, since $D$ is a strictly dominant strategy, defectors will systematically be strictly better off than cooperators: $p_C^t$ will therefore progressively decrease until the total extinction of cooperators. There are two steady states for (7.5), either $(p_C = 1; p_D = 0)$ or $(p_C = 0; p_D = 1)$. The first one is however quite unstable, since any defector who enters the population will outperform the cooperators, whereas the second one is globally asymptotically stable: if $p_D^0 \neq 0$, the population asymptotically converges to $(p_C = 0; p_D = 1)$.
Standard evolutionary theory therefore cannot explain the emergence of cooperative behaviours in PD, since defection remains a strictly dominant strategy, although it would be preferable for each player that both of them cooperate.

### 7.2.2 Cooperation and heuristics

The issue of PD is that $C$ is a strictly dominated strategy, making the socially efficient equilibrium $\{C; C\}$ extremely vulnerable to the invasion of defectors. Consider now a slightly different approach by assuming that individual types are not defined by *actions* but by *heuristics*: successful outcomes, such as mutual cooperation in a PD, would then result from the application of fitted heuristics with respect to the environment in which the individuals choose (Berg and Gigerenzer, 2010). It means that the individuals can implement more subtle strategies and potentially adapt their behaviour according to the nature of the other individuals with whom they are interacting.

A heuristic can be defined as a rule of behaviour that associates to each state of the world the strategy to implement. In standard game theory — as discussed in chapter 6 — the set of "states of the world" corresponds to the set of incomplete strategy profiles of the other players. However, in a strategic interaction, prior to the effective choice of each player, the state of the world cannot be described by the strategies of the player (since they are the *result* of the strategic interaction, not its *premisses*). A game is more adequately described by the *nature* of the players, i.e. of what they intend to do in this game. Their intention can typically be described by a reply function, indicating what the player would do if the strategy of the others were given. In a PD, the set of states of the world based on which the individuals will form their choices is not the set of *strategies* of the players (which are implemented *ex post*), but the set of *reply functions* of the players. In a PD, P1 has therefore beliefs not about P2's *action*, but about P2's *intention*: the specificity of a strategic interaction is precisely that P1 knows that P2 is not simply a machine that follows

a predetermined programme (as in standard evolutionary game theory), but that P2 is a rational agent actuated by certain intentions.
We suggest therefore that each individual is characterised by a reply function (or alternatively by a specific preference relation, from which we could deduce a best reply function). In a strategic interaction, each player then chooses her action according to her own reply function — since it expresses the type of outcomes the player would like to reach — and the reply functions of the others. The *heuristic* of a player is then defined as the mapping that associates to each distribution of reply functions the action to be implemented.
In a PD, the possible reply functions in pure strategies are the following:

- CC: play C whatever the other plays (unilateral cooperator),
- DD: play D whatever the other plays (unilateral defector),
- CD: play C if and only if the other plays C (conditional cooperator),
- DC: play C if and only if the other plays D.

Assume that each player's reply function (her *type*) is common knowledge at each period. Each player intends to play a best reply, knowing that the others are also trying to play a best reply. So as to aggregate those individual intentions, the players must therefore try to reach a strategy profile such that the strategy of each player correspond to the recommendation of their individual reply function. In a strategic interaction, the players are therefore trying to *collectively* reach a strategy profile consistent with the individual intention of each player.

In a PD, the set of strategy profiles consistent with the reply functions of both players can be summarised in the following matrix:

| | CC | CD | DD | DC |
|---|---|---|---|---|
| CC | $\{C;C\}$ | $\{C;C\}$ | $\{C;D\}$ | $\{C;D\}$ |
| CD | $\{C;C\}$ | $\{\{C;C\},\{D;D\}\}$ | $\{D;D\}$ | $\emptyset$ |
| DD | $\{D;C\}$ | $\{D;D\}$ | $\{D;D\}$ | $\{D;C\}$ |
| DC | $\{D;C\}$ | $\emptyset$ | $\{C;D\}$ | $\{\{C;D\},\{D;C\}\}$ |

Since there is an issue of equilibrium existence if one player is a CD-type and the other a DC-type, we assume for sake of simplicity that the type DC is not allowed (this reply function seems indeed dubious in a PD, in the sense that the objective of the player is to reach the asymmetric payoffs, and that $2R > T + S$). Assume

also for convenience that two conditional cooperators always manage to coordinate on the efficient profile $\{C;C\}$. Those cases of equilibrium non-uniqueness will be treated in the next section (the analysis of the PD will then be completed in section 7.4.4).
Under those conditions, we can define the different *heuristics* of player 1 as follows:

- $H(.|CC)$: play C with CC-types, C with CD-types, C with DD-types;
- $H(.|CD)$: play C with CC-types, C with CD-types, D with DD-types;
- $H(.|DD)$: play D with CC-types, D with CD-types, D with DD-types.

The heuristic of a CD-type is therefore to play C with CC and CD-types, but to play D with DD-types. We can then compute the expected payoff of each type:

$$\Pi^t_{CC} = p^t_{DD}S + (1 - p^t_{DD})R, \tag{7.6}$$
$$\Pi^t_{DD} = p^t_{CC}T + (1 - p^t_{CC})P, \tag{7.7}$$
$$\Pi^t_{CD} = p^t_{DD}P + (1 - p^t_{DD})R, \tag{7.8}$$
$$\bar{\Pi}^t = R - (R - P)p^t_{DD}(2 - p^t_{DD}) + p^t_{CC}p^t_{DD}(T + S - 2P). \tag{7.9}$$

Suppose now that individual types are selected according to the material payoff they generate, and their dynamic evolution is represented by (7.5). We have the following result:

**Proposition 9.** *If $p^0_{DD} \neq 0$ and $p^0_{CD} \neq 0$, then:*

$$\lim_{t \to +\infty} p^t_{CD} = 1. \tag{7.10}$$

Proposition 9 means that, whatever the initial distribution of types is[2], it will asymptotically converge to a population composed exclusively of conditional cooperators.

[2]The only restriction on the initial distribution is that there is a non null share of DD and CD-types. Indeed, in the absence of unilateral defectors, everybody systematically cooperate, and the initial distribution of CC and CD remains constant, while the absence of conditional cooperators leads to a situation similar to the one studied in the previous section (the unilateral cooperators get asymptotically extinct).

Our extension of standard evolutionary game theory to types defined by reply function (and therefore indirectly, by heuristics) rather than actions may provide a relevant explanation of the emergence of cooperation in a PD: it indeed appears that this evolution does not only depend on the existence of conditional cooperators, but also on the possibility to sanction defectors. Indeed, the presence of a significant share of unilateral cooperators will be more beneficial to the defectors rather than to the conditional cooperators: it is then necessary to wait until the quasi-extinction of unilateral cooperators by defectors to ensure that conditional cooperators may outperform the defectors — the trajectory is therefore not necessarily monotonic if the initial share of CC is relatively high.
We now would like to argue that this approach in terms of choice of heuristics is probably more accurate than the indirect evolutionary approach.

### 7.2.3 Indirect evolutionary approach

The core idea of the indirect evolutionary approach is that the only way of explaining the emergence of cooperative behaviours in a PD would be to "transform" the payoff matrix such that the players prefer to cooperate when they know that the other cooperates too, i.e. such that the players become conditional cooperators. If a payoff distortion can improve the material payoff of the player — compared to standard PMB and unilateral defection — then such preferences would provide a strategic advantage to those players and a higher fitness. We need therefore to suggest a specific representation of individual preferences, and in particular how we may deduce them from the material payoffs. The first possible approach would be to suggest that some players may present altruistic preferences, i.e. directly cares about the material payoff of the other player (this is for instance the approach of Bester and Güth (1998)). The utility functions of the players would therefore be the following:

$$U_i(x) = \Pi_i(x) + \alpha_i \Pi_{-i}(x), \qquad \forall i \in N. \tag{7.11}$$

Assume that the distorted preferences $U_i$ are common knowledge at each date $t$ (but not necessarily the underlying material payoffs $\Pi_i$). We can easily determine the best reply of player $i$ according to the level of $\alpha_i$:

- (i) $\alpha_i < \frac{T-R}{R-S}$ and $\alpha_i < \frac{T-P}{P-S}$: always play D,
- (ii) $\alpha_i > \frac{T-R}{R-S}$ and $\alpha_i > \frac{T-P}{P-S}$: always play C,

- (iii) $\alpha_i > \frac{T-R}{R-S}$ and $\alpha_i < \frac{T-P}{P-S}$: play C if and only if the other plays C,
- (iv) $\alpha_i < \frac{T-R}{R-S}$ and $\alpha_i > \frac{T-P}{P-S}$: play C if and only if the other plays D.

We can notice that those different best replies precisely define the different reply functions we suggested above: the crucial difference is that it is not certain that all those reply functions are possible. We can indeed notice that the options (iii) and (iv) are incompatible by construction. In particular, if $T-P > R-S$, it is not possible to present preferences such that $i$ is a conditional cooperator (although the procedure itself still makes sense and is evolutionary viable). It means that the particular specification we chose in order to model individual preferences does not allow us to cover every possible reply function: this constitutes a fundamental weakness of the indirect evolutionary approach, since by focusing on altruistic preferences we will be able to explain the emergence of cooperation in a PD if and only if $R-S > T-P$, i.e. when the benefit generated by the cooperation of the other is greater when I cooperate rather than when I defect. This last condition means that the game is supermodular: the benefit of cooperation increases with the number of cooperators. We find here the central assumption of Bester and Güth (1998), who explain the possible invasion of altruistic players within a population of selfish individuals when the game presents strategic complementarities. Bolle (2000) and Possajennikov (2000) highlight on the contrary that relaxing this assumption by considering strategic substitutes leads to malevolent behaviours, i.e. players tend to become more and more aggressive and try to maximise the difference between their payoff and the payoff of others. It seems therefore that being committed to a specific form of payoff distortion may prevent us from identifying evolutionary viable heuristics, for the simple reason that this specific payoff distortion cannot represent the heuristics at stake.

The aim of this illustration was to show that what matters from an evolutionary perspective is not necessarily mere strategies — as in standard evolutionary game theory — or payoff distortions induced by non selfish motives such as altruism or spite — as in the indirect evolutionary approach — but procedures of decision making. Within the context of social dilemmas, it is well-known that reciprocal behaviours may present evolutionary advantages (such as tit-for-tat strategies (Axelrod, 1984)), but it is not certain at all that we may easily represent individual preferences such that their satisfaction implies a reciprocal behaviour: it is probably more relevant to directly focus on the *procedure* of reciprocity rather than trying to find a specification of the utility function such that the corresponding best reply function leads to a behaviour consistent with a reciprocal behaviour.

What drives the evolution of individual preferences within the indirect approach is not a specific payoff distortion, but the commitments it allows and that are beneficial to the players. In a PD, what matters is not being altruistic, but being a conditional cooperator, i.e. being committed to cooperate when the other player cooperates too. By focusing on payoff distortions rather than commitments and reply functions, we may miss possible evolutionary viable strategies, and more importantly lack the fundamental driver of cooperation in social dilemmas, reciprocity.

In this chapter, we provide an analysis of the stability of PMB by studying the dynamics of individual heuristics: each player is characterised by a reply function that determines her heuristic, and — within a population of identical individuals — the share of individuals with more fitted heuristics will progressively grow over time. This approach will also enable us to develop our results with a very few assumptions, completeness and transitivity of the material payoff relation, and perfect observation of the types at each period. We will consider that, rather than maximising a distorted utility function, the players simply interact with each other by following a heuristic. This specification is probably more realistic (in line with the idea of ecological rationality) and better catches the idea of the as-if hypothesis: the individuals are not aware that their actions can be understood as the result of a complex optimisation problem, they are simply following a predetermined rule of behaviour that appeared to be successful in the past. Unlike evolutionary game theory where the players are programmed to play a single strategy, we assume that players are programmed to play a single heuristic. Our approach therefore constitutes a natural extension of evolutionary game theory: we indeed allow for more complex behaviours, while keeping the fundamental idea that the individuals are not maximising a complex utility function but are simply committed to a predetermined pattern of behaviour.

## 7.3 Heuristics game

### 7.3.1 Definitions

Let $N = \{1, \ldots, n\}$ denote the set of players, with $n \geq 2$. $S = \prod_{i \in N} S_i$ denote the set of pure strategy profiles $s = \{s_1; \ldots; s_n\}$, where each set $S_i$ consists of a finite number of pure strategies $s_i$. The *material payoff* of player $i$ is given by the following function:

$$\Pi_i : S \mapsto \mathbb{R}. \tag{7.12}$$

$\Pi_i$ should not be interpreted within this framework as a utility function representing one's true preferences, but rather as a criterion of natural selection (within a biological framework, $\Pi_i$ would for instance define the number of offspring of an individual $i$). We will discuss the possible interpretations of $\Pi_i$ when developing our evolutionary framework.
We assume that player $i$'s type is defined by a *reply function*:

$$R_i : S_{-i} \mapsto S_i. \tag{7.13}$$

$\mathcal{R}_i$ denotes the set of reply functions of player $i$. In particular, we can define the *best reply function* $BR_i \in \mathcal{R}_i$ as the function that associates to any incomplete strategy profile $s_{-i} \in S_{-i}$ the strategy that maximises player $i$'s material payoff:

$$BR_i(s_{-i}) = \arg\max_{s_i \in S_i} \Pi_i(s_i; s_{-i}). \tag{7.14}$$

We need now to define what strategy a player $i$ should play for a given set of reply functions, i.e. to define player $i$'s heuristic, given $i$'s type and the types of the other players. We assume that each player intends to play in accordance with her type, such that each individual strategy of the final strategy profile corresponds to the best reply induced by one's type. The players therefore intend to reach a fixed point of the set-valued mapping of reply functions $R(s) = \prod_{i \in N} R_i(s_{-i})$. If $R(s)$ does not have a unique fixed point, we assume that the players choose a strategy that can be rationalised as the best reply induced by their reply functions given the possible strategies of the other players. So as to define this set of rationalisable strategies, we define the notion of $s$-cycle:

**Definition 8.** *A sequence $C = \{s^k\}_{k \leq K}$ of strategy profiles $s^k \in S$ is a $s$-cycle of $R$ if and only if:*

$$s^k \in R(s^{k-1}), \qquad \forall k \neq 1, \tag{7.15}$$

$$s^1 \in R(s^K). \tag{7.16}$$

A $s$-cycle is a succession of strategy profiles $s^k \in S$ such that player $i$ plays her

best reply $R_i(s^k_{i-1})$ at each step, $\forall i \in N$. After $K$ iterations, the best reply of the players corresponds to the initial strategy profile. If there exists a $s$-cycle $\bar{C}$ such that $K = 1$, then $\bar{s} \in \bar{C}$ is a fixed point of $R$ and is an equilibrium in pure strategies. We can show that:

**Proposition 10.** *$\forall R \in \mathcal{R}$, $R$ has at least one $s$-cycle $C = \{s^k\}_{k \in K}$, $K \geq 1$.*

$\forall i \in N$ and $R \in \mathcal{R}$, we define $C_i(R)$ as the set of pure strategies of player $i$ that belong to a $s$-cycle. Proposition 10 therefore ensures that $C_i(R) \neq \emptyset$. We can now notice that, $\forall i \in N$, the different pure strategies of $C_i(R)$ corresponds to the set of rationalisable pure strategies of the game in normal form $G = \langle N, S, V \rangle$, with $V_i : S \mapsto \mathbb{R}$ the utility function of $i$ such that $i$'s best reply function is $R_i$. Therefore, even in the absence of a pure strategy equilibrium, player $i$ can rationalise the choice of a strategy that is part of a $s$-cycle. We can now define the *heuristic* of player $i$ as follows:

**Definition 9.** *$H_i : \mathcal{R}_{-i} \times \mathcal{R}_i \mapsto S_i$ is the heuristic of player $i$ if and only if:*

$$H_i(R_{-i}|R_i) \in C_i(R_i; R_{-i}), \qquad \forall R_{-i} \in \mathcal{R}_{-i}. \tag{7.17}$$

For a given type $R_i \in \mathcal{R}_i$, the function $H_i(.|R_i)$ associates to each distribution of types among the other players (i.e. to each state of the world) a strategy $s_i$ in $C_i(R)$, i.e. a strategy that can be rationalised as the reply induced by the reply function $R_i$, when the other players intend to play the strategy induced by their respective reply functions $R_j$, $\forall j \neq i$.

We therefore study the game $\Gamma = \langle N, S, \Pi, \mathcal{R} \rangle$, in which the players $i \in N$ choose their strategies $s_i \in S_i$ such that $s_i \in C_i(R)$, and obtain a payoff $\Pi_i(s)$. The objective of this chapter is to determine the conditions under which the players should follow their best reply $BR_i$ so as to maximise their material payoff $\Pi_i$. We now introduce the evolutionary framework.

### 7.3.2 Dynamics

We consider a multipopulation framework in which there exist $n$ large populations of identical individuals in material payoff $\Pi_i$, but with different reply functions. At each date $t \geq 0$, those populations are characterized by a distribution of reply functions $(\theta_1^t; \ldots; \theta_n^t) \in \Delta(\mathcal{R}_1) \times \ldots \times \Delta(\mathcal{R}_n)$, with $\Delta(\mathcal{R}_i)$ the set of probability distributions over $\mathcal{R}_i$. At each period $t$, one individual is randomly drawn from each population and play the game $\Gamma$ defined in section 7.3.1. We assume that the reply function of each player is common knowledge at each period and that $\theta_i^0$ has full support $\forall i \in N$. We do not need to assume that the players know their own material payoff functions $\Pi_i$, i.e. they are not necessarily aware of the criterion of natural selection.
Note that the definition of the criterion of natural selection in economic analysis is not straightforward (see Grüne-Yanoff (2011) on the difficulties of transposing the formal framework of evolutionary game theory from biology to economics), and this is probably why individuals are likely to ignore what $\Pi_i$ is. Although we can reasonably assume that this criterion can be measured by the profits of individual firms in a market (since firms with higher profits are more likely to survive), we can difficultly define an objective standard of "success" in other settings than markets. A possibility would be to refer to the "happiness" criterion, if we accept that what determines the success of an individual is her level of happiness: if a specific heuristic makes me happier (in a mental-state perspective), then I am more likely in the future to follow this heuristic, because I tend to select (unconsciously) the heuristic that made me happy. Another alternative would be to consider the general model of preference evolution we sketched in section 5.2.1: the criterion of evolution would therefore evolve over time according to what the individual perceives as being his self-interest at each date.

This representation of the heuristics game allows us to model a process of natural selection between the reply functions $R_i$ at the individual level: the probability of drawing a player with a specific type increases over time with her average fitness, i.e. the average material payoff she obtains with a reply function $R_i$ given $\theta_{-i}^t$, the distribution of reply functions of other players. So as to properly define a fitness function $V_i : \mathcal{R} \mapsto \mathbb{R}$ — and therefore study the dynamics in the heuristics game — we must treat the possible multiplicity of equilibria in $\Gamma$ for a given set of reply functions $R \in \mathcal{R}$. Indeed, if there exists a player $i$ such that $C_i(R)$ is not a singleton, then several strategies can be rationalised for player $i$, and the strategy of $i$ is not determined.

For each game $\Gamma$, we can define a set of heuristics subgames $\{\Gamma_0^m\}_{m\in M}$ as follows:

$$\Gamma_0^m = \langle N, \mathcal{R}, \{V_i^m\}_{i\in N}\rangle, \tag{7.18}$$

with $\forall i \in N$, $V_i^m : \mathcal{R} \mapsto \mathbb{R}$ a *fitness function* such that:

$$\{V_i^m(R)\}_{i\in N} \in \{\{\Pi_i(s)\}_{i\in N} \mid s \in C(R)\}, \qquad \forall R \in \mathcal{R}, \tag{7.19}$$

$$\bigcup_{m\in M} \{V_i^m(R)\}_{i\in N} = \{\{\Pi_i(s)\}_{i\in N} \mid s \in C(R)\}, \qquad \forall R \in \mathcal{R}. \tag{7.20}$$

The game $\Gamma_0^m$ is therefore a possible heuristics subgame of $\Gamma$, for which a unique vector of material payoff $\{\Pi_i(s)\}_{i\in N}$ has been selected for each profile of reply functions $R \in \mathcal{R}$ (condition (7.19)). The second condition (7.20) means that the set of possible heuristics subgames $\{\Gamma_0^m\}_{m\in M}$ covers every possible combinations of equilibria of $\Gamma$, i.e. that whatever the rules of selection of an equilibrium are in $\Gamma$, there exists a subgame $\Gamma_0^m$ that represents those rules.
Within our framework, each player is programmed to play a given strategy in this heuristics game, and the individuals with more fitted heuristics (generating higher material payoffs in the game $\Gamma$) are more likely to survive. If an individual $i$ can play several strategies $s_i \in C_i(R)$, then the evolutionary process is not deterministic: $i$ can indeed choose any of the strategies from $C_i(R)$, and several trajectories are possible according to the strategy she effectively chooses at each date $t$. We assume that the selection dynamics is monotonically increasing in average fitness, i.e. that distributions of types evolve as follows, $\forall i \in N$:

$$\frac{d}{dt}\theta_i^t(R_i) = g_i(R_i; \theta_{-i}^t), \qquad \forall R_i \in \mathcal{R}_i, \tag{7.21}$$

where $g_i$ is a continuous growth rate functions such that, for any subgame $m \in M$:

$$g_i(R_i, \theta_{-i}^t) > g_i(R_i', \theta_{-i}^t) \Leftrightarrow \int V_i^m(R_i, R_{-i})\,\mathrm{d}\theta_{-i}^t > \int V_i^m(R_i', R_{-i})\,\mathrm{d}\theta_{-i}^t. \tag{7.22}$$

To ensure that $\theta_i$ remains a probability measure for each $t$, we also assume that $g_i$ satisfies, $\forall i \in N$, $\forall t \geq 0$:

$$\int g_i(R_i, \theta_{-i}^t)\,\mathrm{d}\theta_i^t = 0. \tag{7.23}$$

The system (7.21) means that the proportion of individuals with the reply function $R_i$ evolves according to a growth function $g_i$, such that the types with a higher average fitness $\int V_i^m(R_i, R_{-i})\,\mathrm{d}\theta_{-i}^t$ grow faster than types with a lower one (condition (7.22)).

We denote by $\theta^{PMB} \in \Delta(\mathcal{R})$ the distribution of types such that $supp(\theta_i^{PMB}) = \{BR_i\}$, $\forall i \in N$. $\theta^{PMB}$ therefore represents the situation in which every individual from each population $i \in N$ computes her strategy by following her best reply $BR$, i.e. every individual from each population maximises her material payoff. We now introduce our notion of evolutionary stability.

### 7.3.3 Evolutionary stability

A key concept in evolutionary game theory is the notion of evolutionary stability (Maynard Smith and Price, 1973, Maynard Smith, 1982). A strategy is evolutionary stable if it performs strictly better than any "mutant", as soon as the initial invasion of mutants is small enough. The interesting property of this notion of stability is that an evolutionary stable strategy is asymptotically stable in the replicator dynamics (Taylor and Jonker, 1978): although the population may be invaded by a low share of mutants, the evolutionary dynamics will asymptotically restore the initial population. This notion of stability is however quite strong, in particular in a multipopulation framework, in which a strategy profile is evolutionary stable if and only if it is a strict Nash equilibrium (Weibull, 1995, p.167). In the case of the heuristics game, we can notice for instance that there almost never exists an evolutionary stable strategy in the sense of Maynard Smith and Price (1973): it is indeed possible to change one's reply function without impacting the equilibrium of the game. The mutants will therefore perform as well as the other players and will not be eliminated by evolutionary pressures (in this situation, players will behave *as if* they were maximising their material payoff).
The notion of neutral stability has then been developed as a weaker stability concept, without requiring the strictness property of evolutionary stability: although the replicator dynamics will not restore the initial state, the population of mutants will not grow after the invasion — they are indeed unable to outperform the individuals from the initial distribution. A question that arises then is to know whether the invasion of neutral mutants could generate more instability, by creating the opportunity for mutants to enter and successfully invade another population. As an illustration of this phenomenon, consider the case of the PD studied above. Assume that the two populations from which players 1 and 2 are

drawn are composed exclusively of DD-types[3]. If population 1 is invaded by a fraction $\varepsilon$ of CD-types, then those mutants will also defect (no one is indeed likely to cooperate in the second population). Being a unilateral defector with probability 1 is therefore a neutrally stable strategy. However, now that a few CD-types invaded population 1 (notice that they cannot be outperformed by DD-types, since everyone is defecting), there is an opportunity for CD-types to invade population 2. Proposition 9 states that the DD-types will then progressively disappear to the benefit of CD-types. Although PMB is neutrally stable — it is indeed immune against the invasion of a single population — it is not immune against the invasion of several populations (either simultaneously or successively): the invasion of a neutral mutant in population 1 gives the opportunity to the invasion of mutants in population 2 — invasion that would not have been successful if population 1 had not been invaded previously by CD-types.
van Veelen (2012) suggests the notion of robustness against indirect invasion (RAII) to embody this idea: although some mutants may enter the population and perform as well as the initial population, they do not give an opportunity for new mutants to successfully invade the new population. It is shown in particular that RAII strategies (and their neutral mutants) are asymptotically stable in the replicator dynamics.

Since our objective is to test the as-if hypothesis, we should offer a notion of evolutionary stability that tolerates the invasion of neutral mutants who behaves as if they were payoff maximisers (i.e. such that the fixed point of the heuristics mapping is a Nash equilibrium), but not of neutral mutants whose heuristics leads to an equilibrium that is not a Nash equilibrium. We therefore firstly define a notion of "as-if evolutionary stability":

**Definition 10.** *$\bar{\theta} \in \Delta(\mathcal{R})$ is an as-if evolutionary stable strategy if and only if there exists a subgame $\Gamma_0^m$ such that, $\forall i \in N$, $\forall R_i \in supp(\bar{\theta}_i)$:*

[3]Our discussion of the PD in section 7.2 assumed that both players were drawn from the same population: we consider now a multipopulation framework in which each player is drawn from a specific population. Note that proposition 9 would remain true in a multipopulation framework: as long as there exists DD and CD types in both populations, they asymptotically converge to a distribution of types composed exclusively of CD-types.

$$\int V_i^m(R_i; R_{-i})\,\mathrm{d}\bar{\theta}_{-i} = \int V_i^m(R_i'; R_{-i})\,\mathrm{d}\bar{\theta}_{-i}, \qquad \forall R_i' \in supp(\bar{\theta}_i), \tag{7.24}$$

$$\int V_i^m(R_i; R_{-i})\,\mathrm{d}\bar{\theta}_{-i} > \int V_i^m(R_i''; R_{-i})\,\mathrm{d}\bar{\theta}_{-i}, \qquad \forall R_i'' \in \left\{\mathcal{R}_i \;\middle|\; C(R_i''; \bar{R}_{-i}) \neq C(\bar{R}), \bar{R} \in supp(\bar{\theta})\right\}. \tag{7.25}$$

Condition (7.24) means that $\bar{\theta}$ is a steady state of the replicator dynamics (all the different types within the population $\bar{\theta}$ obtain the same average expected payoff) and condition (7.25) that the different types within the population at this steady state always achieve strictly more than what a mutant $R_i''$ would achieve by invading the population $i$, when the equilibrium played by the postentry population is different from the equilibrium in the initial population.
An as-if evolutionary stable strategy is therefore immune against the invasion of a *single* population by a mutant $R_i''$: it is however not sure whether the invasion of population $i$ by a neutral mutant can generate a potential gain for the invasion of a new mutant in a population $j \neq i$. Such a notion of stability against the invasion of *several* populations is relatively similar to the notion of robustness against indirect invasion developed by van Veelen (2012)[4]. We define the set $N_i(\bar{\theta})$ of neutral mutants for $i$ as the set of reply functions $R_i'' \notin supp(\bar{\theta}_i)$ such that:

$$\int V_i^m(R_i; R_{-i})\,\mathrm{d}\bar{\theta}_{-i} = \int V_i^m(R_i''; R_{-i})\,\mathrm{d}\bar{\theta}_{-i}, \qquad \forall R_i \in supp(\bar{\theta}_i). \tag{7.26}$$

If the initial distribution is invaded by players with reply functions from $N_i(\bar{\theta})$, then those players will perform as well as the players in the initial population. Random shocks may increase the share of those neutral mutants within the population $i$ without being eliminated by evolutionary pressures. We can thus drift to a distribution $\hat{\theta} = \alpha\bar{\theta} + (1-\alpha)\tilde{\theta}$, with $supp(\tilde{\theta}) \subseteq N_i(\bar{\theta})$ and $\alpha \in [0;1]$. Although we know by construction that the condition of as-if evolutionary stability is still verified within population $i$, it is not certain any more that it is still the case in other populations $j \neq i$: the invasion of neutral mutants in population $i$ may generate an opportunity for new mutants to enter in other populations. The notion of robustness against indirect invasion catches the idea that, although the initial distribution can be invaded by neutral mutants, the new distribution will remain as-if evolutionary stable:

[4]We must nevertheless adapt Van Veelen's notion to a multipopulation framework.

**Definition 11.** *$\bar{\theta} \in \Delta(\mathcal{R})$ is robust against indirect invasion if and only if, for all sequences $\{\theta^k\}_{k\in K}$ such that:*

- $\theta_i^0 = \bar{\theta}_i$,
- $\theta_i^{k+1} = \alpha_i^k \theta_i^k + (1-\alpha_i^k)\tilde{\theta}_i^k$, *with* $supp(\tilde{\theta}_i^k) \subseteq N_i(\theta^k)$ *and* $\alpha_i^k \in [0;1]$, $\forall i \in N$, $k \in K$,

*$\theta^k$ is as-if evolutionary stable $\forall k \in K$.*

A distribution $\bar{\theta}$ is robust against indirect invasion if the invasion of the initial populations by neutral mutants (and then of the resulting distribution by their own neutral mutants) does not generate more instability, i.e. the distribution $\theta^k$ remains as-if evolutionary stable over time after multiple invasions by neutral mutants. This notion of stability describes a situation of stability against the successive invasion of several populations.

## 7.4 Stability of payoff-maximising behaviour

We can now study whether PMB is evolutionary stable in the replicator dynamics. We will then be able to validate or invalidate the as-if hypothesis, according to which players manage to maximise their material payoff, although it is not necessarily their intention.

### 7.4.1 $\phi$-core

Before stating our main result, we introduce the notion of the $\phi$-core. Suppose that the game $\Gamma$ is played cooperatively, but without transfer among the members of a coalition. Several notions in cooperative game theory have been developed so as to specify for each coalition the payoff that can be ensured to its members in $\Gamma$, according to the behaviours of the players outside the coalition $S \subset N$. The $\gamma$-core (Chander and Tulkens, 1997) corresponds to the set of vectors of payoffs such that no coalition $S \subset N$ has an interest in breaking the agreement, and play the underlying normal form game as a coalition against the set of remaining players $N \setminus S$, each one acting as singletons and maximising her own payoff. Currarini and Marini (1998) suggest an alternative notion, by arguing that the formation of the coalition and the choice of a coordinated strategy in the underlying normal form

game are simultaneous events (unlike the $\gamma$-approach that relies on a two stage game structure: a coalition forms in the first stage, and then plays against the other players, split up as singletons in the second stage). The deviating coalition therefore possesses a first mover advantage, by taking into account the reaction of the excluded players. Our aim is not to discuss the relevance of the $\phi$-core for cooperative game theory, but this notion appears to be perfectly suited to our analysis.
For two vectors of payoff $\Pi_M(s) = \{\Pi_i(s)\}_{i\in M}$ and $\Pi_M(s') = \{\Pi_i(s')\}_{i\in M}$, we write $\Pi_M(s) > \Pi_M(s')$ if $\Pi_i(s) \geq \Pi_i(s')\ \forall i \in M$, with at least one strict inequality.

**Definition 12.** *$\forall M \subseteq N$, let $St(M) \subset S$ denote the set of Stackelberg equilibria such that the coalition $M$ is a Stackelberg leader and $N \setminus M$ its followers:*

$$\begin{aligned} St(M) = \{(\tilde{s}_M; BR_{-M}(\tilde{s}_M)) \in S \mid \nexists s_M \in S_M, \\ s.t.\ \Pi_M(s_M; BR_{-M}(s_M)) > \Pi_M(\tilde{s}_M; BR_{-M}(\tilde{s}_M))\}, \end{aligned} \tag{7.27}$$

$\tilde{s} \in St(M)$ is a Stackelberg equilibrium of the coalition $M$ if and only $M$ cannot Pareto improve the utility of all its members simultaneously, knowing the best reply of the other players. We define the $\phi$-core as follows:

**Definition 13.** *Let $\Gamma = \langle N, S, \Pi, \mathcal{R}\rangle$ be a game as defined in 7.3.1. The $\phi$-core of $\Gamma$ is the set of vectors of payoffs such that no coalition $M \subseteq N$ can improve the payoffs of all its members:*

$$\phi - core = \{\{\Pi_i(s)\}_{i\in N} \mid \forall M \subseteq N, \nexists \tilde{s} \in St(M)\ s.t.\ \Pi_M(\tilde{s}) > \Pi_M(s)\}. \tag{7.28}$$

It follows from this definition that all the elements of the $\phi$-core are Pareto optimums: otherwise, the grand coalition could deviate to a Pareto superior outcome. We can for instance check that the $\phi$-core of the prisoner's dilemma presented is section 7.2 is non empty, with:

$$\phi - \text{core} = \{(R; R)\}. \tag{7.29}$$

If one player decides to break the agreement and plays as a Stackelberg leader,

then she will only get $P$, since defecting is a strictly dominant strategy for her follower. Furthermore, it is not possible to collectively deviate to another strategy profile such that both players will be better off. $(R;R)$ is the only element of the $\phi$-core since mutual defection is Pareto dominated and is not immune against a collective deviation, and the asymmetric payoffs do not ensure to all the players their Stackelberg payoff $P$: the cooperative player would indeed unilaterally deviate from an asymmetric agreement and play the game as a Stackelberg leader.

A possible limitation of the $\phi$-core is that some coalition $M \subset N$ can still be indifferent between the vector of the $\phi$-core and a Stackelberg equilibrium in which $M$ is leader. In particular, although $M$ has no incentive to deviate from the agreement, it has also no incentive to accept the agreement. We therefore define the *strict $\phi$-core* as a refinement of the $\phi$-core by considering only the vectors of payoff such that no coalition can achieve on its own at least the payoff of the core:

**Definition 14.** *Let $\Gamma = \langle N, S, \Pi, \mathcal{R} \rangle$ be a game as defined in 7.3.1:*

$$\textit{strict } \phi\textit{-core} = \{\tilde{s} \in \phi - \textit{core} \mid \forall M \subset N, \nexists \tilde{s} \in St(M) \textit{ s.t. } \Pi_M(\tilde{s}) \geq \Pi_M(s)\}. \tag{7.30}$$

### 7.4.2 Evolutionary stability of PMB

We can now state our main result:

**Proposition 11.** *Let $\Gamma = \langle N, X, \Pi, \mathcal{R} \rangle$ be a game as defined in section 7.3.1, and $\bar{s} \in C(BR)$:*

*(i) $\theta^{PMB}$ is as-if evolutionary stable if and only if, $\forall i \in N$, $\nexists \tilde{s} \in St(i)$, $\tilde{s} \neq \bar{s}$, s.t. $\Pi_i(\tilde{s}) \geq \Pi_i(\bar{s})$;*

*(ii) $\theta^{PMB}$ is robust against indirect invasion if and only if $\bar{s}$ is in the strict $\phi-$core.*

Proposition 11 means that PMB is as-if evolutionary stable if and only if $\bar{s}$ is the unique Stackelberg equilibrium of the game, whoever the leader is (claim (i)).

No player can therefore benefit from a first mover advantage, since $i$, when she is a Stackelberg leader, cannot reach an equilibrium $\tilde{s} \neq \bar{s}$ that gives her at least her Nash payoff, $\forall i \in N$. If at least one player $i$ does not achieve her Stackelberg payoff at Nash equilibrium, then the population $i$ can be successfully invaded by mutants whose reply function is not their best reply function $BR_i$. Claim (ii) states that PMB is robust against indirect invasion if and only if no coalition $M \subseteq N$ can benefit from a first mover advantage: the Nash equilibrium $\bar{s}$ must therefore ensures to all the possible coalitions their Stackelberg payoff. PMB is therefore stable against the simultaneous invasion of several populations if and only if the resulting Nash equilibrium $\bar{s}$ is in the strict $\phi$-core (no mutants should indeed be allowed to invade the population and reach a Stackelberg equilibrium different from the Nash equilibrium, without being outperformed by payoff maximisers). Proposition 11 can be seen as a generalisation of the proposition 1 of chapter 5, since we are studying the set of games for which a player could benefit from adopting a heuristic distinct from payoff maximisation (by choosing a type $R_i \neq BR_i$): the main difference between those two results is that our analysis in chapter 5 only considered preference distortions that could be expressed as a weighted sum of the material payoffs, whereas we are directly studying the choice of reply functions in this chapter.

We can notice that the set of games such that those conditions are met is quite restricted: we can mention for instance zero-sum games (for which there is no first mover advantage) or games presenting a strategy profile that Pareto dominates all the other outcomes (a Stag Hunt for instance). An interesting implication of our result is that PMB will be evolutionary stable if and only if players select the payoff-dominant equilibrium in presence of multiple Nash equilibria (Harsanyi and Selten, 1988). This result contrasts with the results of Kandori et al. (1993), who show that, in symmetric $2 \times 2$ games with two symmetric strict Nash equilibria and random mutations, the evolutionary dynamics will select the payoff-dominant equilibrium if and only if the strategies have equal security levels. Studying the evolution of heuristics rather than the evolution of individual strategies may therefore provide a rationale for the selection of payoff-dominant but risky equilibria. A crucial assumption allowing for this result is however the perfect observability of one's heuristics: relaxing this assumption (by considering for instance a noisy perception of the heuristic of the other players) would probably privilege safer equilibria.

We may also notice that proposition 11 enables us to generalise the results of Bester and Güth (1998), Bolle (2000) and Possajennikov (2000) concerning the connection between supermodularity and the evolutionary sustainability of cooperative

motives. We indeed show that the general dynamics of preference evolution is driven by the possibility for some players to benefit from a strategic commitment, which will be the case as soon as they may benefit from a first mover advantage. In particular, in case of positive best replies, being a Stackelberg leader is equivalent to being more cooperative: in supermodular games, since evolutionary pressures will select players who behave like Stackelberg leaders, cooperative motives will be evolutionary stable. On the contrary, submodularity will select aggressive players, since such motives allow them to behave as if they were Stackelberg leaders.

### 7.4.3 Team reasoning as an ecologically rational heuristic

Proposition 11 provides necessary and sufficient conditions under which PMB is evolutionary stable. A complementary perspective would be to determine the nature of the heuristics that are likely to outperform PMB when the Nash equilibrium is not in the strict $\phi$-core. In the proof of proposition 11, we introduce the type $\tilde{R}$:

- $\forall i \in M$, $\tilde{R}_i(s_{-i}) = BR_i(s_{-i})$, if $\exists j \in M$ s.t. $s_j \neq \tilde{s}_j$,
- $\forall i \in M$, $\tilde{R}_i(s_{-i}) = \tilde{s}_i$, if $\forall j \in M$, $s_j = \tilde{s}_j$,

with $\tilde{s} \in St(M)$, and $\Pi_M(\tilde{s}) > \Pi_M(\bar{s})$. $\tilde{R}_i$ can be described as following: "maximise your material payoff in every circumstances, unless all the players of $M$ play the optimal strategy of the coalition". $\tilde{R}$ captures the idea that players in the coalition $M$ are reciprocators: they accept to play their part of the optimal collective strategy profile if and only if the others do their part of the work. In particular, since there may be an issue of equilibrium selection if all the players $j \in M$ effectively engage in strategic interaction with the reply function $\tilde{R}_j$, the only way to ensure that the heuristic $H_j(.|\tilde{R}_j)$ will systematically outperform PMB is that, when all the players meet and actually identify themselves as pursuing the same collective objective (achieving the Stackelberg equilibrium of their coalition), they select the dominant strategy profile for the team. The reply function $\tilde{R}$ and the possibility to select the good equilibrium when all the members of $j$ intend to reach the Stackelberg equilibrium of the team precisely embodies the idea that the players $j \in M$ are team reasoning. This means that team reasoning can be interpreted as following a heuristic when one's type is $\tilde{R}$: team reasoning is therefore an ecologically rational heuristic, since being able to team reason is more fitted than PMB to the game within which players interact.

If personal intentions are observed, then it is my interest to declare myself as a team reasoner, since this may generate the opportunity for other individuals to

declare themselves as team reasoners too, and then to achieve together a collectively desirable outcome. Players of type $\tilde{R}$, and who select the Stackelberg equilibrium of the team when all the players $j \in M$ are also of type $\tilde{R}$ therefore behave as if they were team reasoning.

More generally, we can show the following proposition:

**Proposition 12.** *Let $\bar{\theta}$ be a distribution of types such that, $\forall i \in N$, $supp(\bar{\theta}_i) = \{\bar{R}_i\}$ and $C_i(\bar{R}) = \{\bar{s}_i\}$. Let $\hat{s}$ be a strategy profile Pareto superior to $\bar{s}$. If $\bar{\theta}$ is as-if evolutionary stable, and $\hat{\theta}$ is such that:*

- $supp(\hat{\theta}_i) = \{\hat{R}_i\}$,
- $\hat{R}_i(s_{-i}) = \bar{R}_i(s_{-i})$ *if* $s_{-i} \neq \hat{s}_{-i}$,
- $\hat{R}_i(\hat{s}_{-i}) = \hat{s}_i$,

*then $\hat{\theta}$ is as-if evolutionary stable.*

Proposition 12 means that, if there exists a distribution of profiles that is as-if evolutionary stable[5], then we can construct a distribution of profiles that is also as-if evolutionary stable, and in which all the players are team reasoning: their type indeed induces them to play the collectively rational profile $\hat{s}$ if and only if all the other players have the same type. This means that, as soon as there exists an evolutionary stable distribution of type, there also exists an evolutionary stable profile with team reasoners. Furthermore, the team reasoners gain at least what the players of the initial distribution gains: team reasoners are therefore more likely to be robust against indirect invasions, since their distribution is already protected by construction against the invasion of the grand coalition.

[5]Notice that we have restricted the proposition to distributions in which we do not have to deal with multiple equilibria: proposition 12 can however be extended to include those cases, since the underlying logic will be exactly the same: the type $\hat{R}_i$ would be defined such that all the players play the collectively rational profile $\hat{s}$ when all the players have the type $\hat{R}$, while the interaction of other types will lead to the same outcome than with the initial distribution.

### 7.4.4 Illustration

Consider now the PD discussed above: in section 7.2, we explicitly excluded the possible cases for which there was not a unique equilibrium for a given set of reply functions. We can now also treat the cases of equilibrium non-uniqueness. Recall that 4 reply functions in pure strategies are available to the players, CC, DD, CD and DC. The problematic cases concerns the interaction of CD and DC types: when two CD or two DC met, there exist two pure strategy Nash equilibria, whereas the interaction of a CD type with a DC type does not have any pure strategy equilibrium. In all cases we can however notice that we have for both players $C_i(R) = \{C; D\}$, i.e. that they can all rationalise the choice of one of the two available pure strategies. The general payoff matrix for each heuristics subgame is the following:

| | CC | DD | CD | DC |
|---|---|---|---|---|
| CC | $(R;R)$ | $(S;T)$ | $(R;R)$ | $(S;T)$ |
| DD | $(T;S)$ | $(P;P)$ | $(P;P)$ | $(T;S)$ |
| CD | $(R;R)$ | $(P;P)$ | $\Pi(CD;CD)$ | $\Pi(CD;DC)$ |
| DC | $(T;S)$ | $(S;T)$ | $\Pi(DC;CD)$ | $\Pi(DC;DC)$ |

the payoffs in the last 4 cells are not determinate and depend on the strategy chosen by the players from each population (the 4 vectors of payoff are possible, since both strategies are rationalisable for both players). There exist $4^4 = 256$ possible heuristics subgames, according to whether each type CD and DC from each population cooperates or defects with CD and DC types from the other population. We can then notice that, in the different subgames, only two pure strategy Nash equilibria are possible: $\{DD; DD\}$ (for any subgame) and $\{CD; CD\}$ (in the subgames in which CD cooperates with CD, and either DC cooperates with CD or CD defects with DC).

PMB is therefore as-if evolutionary stable in the replicator dynamics described in (7.21): $D$ being a dominant strategy for both players, the Nash equilibrium is also a Stackelberg equilibrium of both players. We can now notice that an invasion of CD in a population of DD will achieve a higher average payoff (as described in section 7.2) if and only if CD types manage to coordinate on the efficient equilibrium $\{C; C\}$. Payoff maximisers can therefore be invaded by a population of mutants whose reply functions are defined in the proof of proposition 11, i.e. who play the optimal strategy of the coalition $N$ if and only if the other members of the coalition play their part of this optimal strategy profile (conditional cooperation): this means that maximising one's payoff is not robust against indirect invasion, since the entry of CD types within a population of DD for population 1 is neutral only if CD types

do not invade also the second population.

## 7.5 Conclusion

The aim of this chapter was to establish necessary and sufficient conditions under which PMB is evolutionary stable: we showed that apart from very specific games such that zero-sum games or games with a Pareto dominant outcome, players are likely to adopt apparently irrational patterns of behaviours, by computing their strategy without referring to their best reply function. We showed that when the players adopted heuristics outperforming payoff maximisers, they were behaving as if they were team reasoning: they indeed adopt reciprocal heuristics and are able to coordinate on the collectively desirable strategy profile[6]. We suggested amending the indirect evolutionary approach by studying the evolution of individual heuristics rather than the evolution of payoff distortions, since it appears that modelling the evolution of preferences requires determining *a priori* a general form of utility function, which cannot necessarily represent any possible preference distortion (since it is restricted to the payoff distortions induced by the specification of the utility as a function of the material payoff functions). It is therefore more relevant to directly study the evolution of individual heuristics and commitments, since it is fundamentally those elements that drive the evolution of individual preferences within the indirect evolutionary approach.

Our main theorem is fairly intuitive, since we show that as soon as a player (or more generally a coalition) may benefit from a strategic commitment, PMB is not evolutionary stable: it will indeed be in the interest of this player to behave as if she was a Stackelberg leader. Furthermore, although we restricted our analysis to finite games, the reasoning still remains valid for games with more complex set of strategies: the only issue is to precisely define what is the acceptable set of strategies for a player knowing the distribution of reply functions $R$ (in particular if there is no Nash equilibrium, even in mixed strategies). Our result also enabled us to generalise some results of the indirect approach literature, and in particular the tight connection between supermodularity and the evolutionary stability of cooperative behaviours. Being a Stackelberg leader in a supermodular game is indeed strategically equivalent to being more cooperative: evolutionary pressures selecting individuals behaving like Stackelberg leaders, it will therefore sustain cooperation in supermodular games and competition in submodular games.

[6] We can also interpret the individual deviations from PMB as a form of team reasoning (as in section 6.5), since players are choosing the collective preferences that maximises the payoff of the team reasoners, i.e. their own payoff.

Our work can be extended in several directions: the crucial assumption of this chapter is that types are perfectly observed at each period. In particular, since a strategic commitment is efficient if and only if the other players know it, we can wonder whether the uncertainty about the other's types can still generate an opportunity to behave as a Stackelberg leader, in particular if the learning process of the different players about the types of each other is relatively slow. In this situation, a player committed to play a Stackelberg strategy will indeed achieve low payoffs as long as the others do not recognise her type. Another development would be to study the conditions under which the players effectively converge to PMB (and not merely whether PMB is evolutionary stable). This may also provide some evolutionary arguments in favour of specific rules of equilibrium selection: we indeed showed here that payoff maximisation is more likely to survive if the players refer to a payoff dominance criterion rather than (for instance) a criterion of risk dominance, since this could lead to a Nash equilibrium outside the strict $\phi$-core.

# Conclusion

By questioning the hypothesis that individual preferences are consistent and context-independent, behavioural economics also questions the validity of the traditional preference-satisfaction criterion in normative economics. Our first objective in this thesis was to question the methodological accuracy of behavioural welfare economics, according to which (i) individuals are defined by true preferences — similar to the ones we find in neoclassical economics — that would determine their choices if they were able to reason correctly, and (ii) the satisfaction of those true preferences is the normative criterion. The difficulty of this model is that it requires the existence of a latent mode of reasoning able to generate counterfactual coherent preferences (chapter 1): we however suggested that the assumption of coherent preferences was a property of the representative agents interacting in repeated markets, and does not necessarily accurately describe the actual behaviour and preferences of the real agents interacting in markets (chapter 2). We indeed suggested that the idea of latent rational preferences was already implicit in neoclassical economics from Pareto on, and that behavioural welfare economics could be seen in the direct continuation of Pareto's reductionist analysis of individual behaviour (chapter 3). Furthermore, although satisfying one's true preferences — if such preferences may exist — seems to be a reasonable normative criterion, we argued that this approach fundamentally denies individual autonomy, and sees the individuals as passive utility producers, unable to form normative judgements and to lead their own life. We argued instead that the normative issue raised by behavioural findings is that people may lack of autonomy: their choices are indeed likely to be manipulated by third parties without their consent, and boundedly rational individuals may not be able to choose the life they have reasons to value, and endorse the full responsibility of their choices (chapter 4).

It seems therefore that BWE may not provide an adequate solution to the reconciliation problem: our interpretation of behavioural findings is not that economic agents are faulty Econs (as implicitly assumed in BWE) but rather that economic agents are Humans, for whom it does not make sense to determine true preferences, whose satisfaction matters. Real agents "carry around with them a rag-tag bundle of values, desires, taboos, moral commitments, and wishes along with a few firmly held preferences and, when asked to choose, they generate on the fly a ranking of the alternatives under consideration" (Hausman, 2015), but nothing justifies that, if real agents were able to reason perfectly (as an Econ) instead of using complex psychological processes, they would generate counterfactual context-independent

preferences. Reconciling normative and behavioural economics therefore seems to require providing an alternative model of preferences, that does not rely on a notion of *true* preferences.

We stressed in chapter 4 that behavioural welfare economics, in the direct continuation of neoclassical welfare economics, considers the individual as a passive agent, whose only objective is the satisfaction of her exogenous preferences. The issue with this picture of the individual is that it lacks a notion of reflexive agent, i.e. an entity "that can examine one's values and objectives and choose in the light of those values and objectives" (Sen, 2002, p.36). Hollis (1998) for instance argues that "we need a more social conception of what persons are and a role-related account of the obligations which make the social world go round and express our humanity" (p.104). We therefore suggested developing in the second part of this thesis a model of endogenous preferences, in which individuals are able to build their own preferences and identity through their multiple commitments to groups of players they interact with. Furthermore, by grounding our model on Bacharach's variable frame theory, we were able to get rid of a notion of true preferences as the primitive of our model. The rankings of the outcome under the I-frame are indeed better understood as a matter of psychological salience: the I-frame characterises my psychological perception of the states of the world (which can therefore be informed by experimental psychology), and players are then able to choose their own preferences, given their initial perception of the game. Our model of team reasoning, by stressing the central role of the construction of collective preferences (interpreted as collective commitments), may therefore offer a more satisfying model of individual behaviour. We argued that team reasoners are able to choose their collective preferences and determined the optimal collective preferences each team should choose: we highlighted in particular that team reasoners are likely to choose aggressive preferences with outsiders in games with strategic substitutes, while they will tend to adopt cooperative preferences with outsiders in games with strategic complementarities (chapters 5 and 6). We then showed that rational players should be able to choose their own frames, and that it is generally in their interest to team reason (chapter 6). We finally showed that actual players are likely to become team reasoners, since team reasoning can be interpreted as an ecologically rational heuristic (chapter 7).

## Team reasoning and the reconciliation problem

We would like to discuss now how the model of preferences we developed in the second part of this thesis may contribute to solve the reconciliation problem. We already highlighted that, contrary to the model of the inner rational agent of BWE, our model does not rely on the existence of true preferences whose satisfaction is normatively desirable. Furthermore, it explicitly provides a picture of the individual as a socially embedded agent, whose identity depends on the multiple individuals with whom she interacts. A natural extension of our model (sketched by Bacharach in his 2006 book) would be to consider that individuals can also be considered as a "team of selves", rather than a simple collection of selves who may have contradictory interests. An individual like Oscar is then more than a set of transient selves $t$-Oscar: he has his own existence as an enduring agent (whose existence is ensured by the continuity of agency of the temporal selves $t$-Oscar at each date $t$), and as the supervisor of his team of selves. Oscar can then define a course of actions that the different selves should follow: this allows him to form commitments, such that each $t$-Oscar does not consider himself as an independent individual, but as part of a greater plan involving his "fellow selves". In this situation, what matters for the decision maker at date $t$ is not the satisfaction of the preferences of $t$-Oscar, but the satisfaction of the preferences that Oscar, as an autonomous agent, has chosen knowing the possible inclinations of his future selves and their likelihood to respect or not his plan.

This description of the individual may for instance embody Sugden's notion of a *responsible agent*: each self is responsible for her choices, as well as for the choices of her past and future selves (although those choices may be motivated by preferences that the present self does not recognise), because the identity of the agent is unified by a continuous *locus* of responsibility. Sugden (2004, p.1018) argues that this notion of responsibility provides a philosophical justification for the claim that opportunity has value (Cohen, 1989, Sen, 1998, Arrow, 1995, Roemer, 1998). According to the opportunity criterion, the fact that retired-Oscar regrets the choice of young-Oscar cannot justify a paternalistic intervention on young-Oscar, since retired-Oscar, as part of the continuing and responsible agent Oscar, is responsible for the choices of his past selves. Our position is however slightly different, since we argue that what matters is not opportunity, but the satisfaction of the preferences that Oscar, as an autonomous agent, has chosen. As an illustration of our position, consider the choice of SuperReasoner in Sunstein and Thaler cafeteria. Suppose that SuperReasoner's preference between the fruit and the cake is indeterminate, i.e. he has no decisive reason for preferring one item to the other. SuperReasoner

must therefore *choose* his own preferences: what matters here is not the satisfaction of some exogenously given preferences, but the satisfaction of *his own* preferences, of the preferences he has chosen.
In this perspective, what behavioural economics tells us is not that real individuals are poor decision makers, but that they lack of autonomy, since they are likely to ruin their long-term plans due to irrelevant framing effects that affect their transient selves. The "own good" of Oscar can only be defined by Oscar himself, as an autonomous and enduring agent: although it seems that behavioural findings may call for paternalistic measures, they cannot consist in imposing to Oscar what we think he would have preferred if he were autonomous. The normative value of the satisfaction of Oscar's preferences indeed lies in the fact that Oscar has freely chosen them. Therefore, even if the choice architect manages to discover the preferences that Oscar would have chosen (although this is very unlikely, since Oscar's freedom of choice precisely implies to some extent a form of unpredictability of his choice), satisfying those preferences is not normatively desirable: they are indeed not *his own* preferences, since he has not chosen them. Paternalistic interventions should only consist in helping Oscar to be more autonomous and to give him the opportunity to choose what matters for himself, i.e. to choose the life he wants to live. Possible measures in this direction would typically consist in educating the individuals by warning them of the existence of framing effects, or more generally of the diverse socio-psychological biases that are likely to affect their choices. A complementary set of measures would then consist in providing to the individuals the means to make and enforce their commitments (such as giving to the users of a CPR the means to organise their own institutional rules). Consider for instance the case of Patrick, discussed in chapter 4: due to his relatively modest social origins, he does not even consider the option of going to university. By educating him, and trying to convince him that his low social aspiration (and his self-censorship) is the direct consequence of his modest origins (and that Rachel, with the same academic abilities but wealthier origins, will pursue her studies), Patrick could re-evaluate his choice. The government could then financially help Patrick to pursue his studies (his modest origins indeed also implies that he may have important funding constraints).

## Future research

The results of this thesis can be extended in several directions. Firstly, we could investigate possible measures of the level of "individual autonomy", so as to be able to compare different alternatives according to the degree with which the individuals

have the ability of choosing their own preferences (this may for instance consist in a first step in developing a measure of the set of "psychologically possible" options). For instance, we suggested that a central element of individual autonomy was the ability to choose one's collective preferences: we must therefore ensure the proper functioning of democratic processes, such that the individuals may collectively and publicly debate, and define *in fine* what matters for them, as a group.
Another dimension that could be explored — and which has been ignored in the present work — is the evaluation of non-market goods thanks to contingent evaluations. Since individual preferences are very volatile and does not necessarily tell us something about individual well-being, does it still make sense to make cost-benefit analyses based on such evaluations? Suppose for instance that the government tries to evaluate the value of the conservation of polar bears. If you are asked the amount of money you would be ready to pay to ensure that polar bears do not get extinct, can we simply evaluate the value of conservation by aggregating the willingness to pay of each individual? Our recommendation would probably be in line with the argument we developed in chapter 4: instead of seeing in our preferences (or more precisely, in the preferences that the economists would purify from what they think are errors) the only source of normativity for making social evaluations, we should simply facilitate the public debate on this issue so as to decide collectively whether the conservation of polar bears matters or not. An important feature of this approach is that it does not let normative economists in charge of the evaluation (although it allows the citizens to choose to delegate this question to economists), and therefore promotes a democratic rather than technocratic approach to tackle social issues.

The other direct extensions of our results concern our model of preferences. The first line of research would be to define more precisely a measure of salience, i.e. how individuals tend to perceive the outcomes of their choice problem. The issue of defining I-preferences is that we cannot simply state that this perception is revealed by the choices of the individual: my final choice indeed depends on the perception of the game under my I-frame, *and* of my intentions (either individual or collective). Observing cooperative behaviours in a PD does not necessarily mean that the psychological perception of the game is compatible with inequity aversion, since it can also be the result of a "material" perception of the game (in terms of individual monetary gains) and of the expression of collective intentions. A possibility to test experimentally this difference between perception under the I-frame and intentions would be to observe the different attitudes of subjects when they are told that they play a game against another human or against a computer

with a predetermined strategy: while it is possible to identify with the other human in the former case (and therefore to form collective intentions, because reciprocity is possible), the subject is playing *against* a computer in the latter, and cannot expect any reciprocity from a predetermined computer.

Several extensions of our theoretical framework of team reasoning are possible. The first development could be the study of the mode of aggregation of preferences within a given team (we indeed assume in chapter 6 that team reasoners use the Nash solution to select their collective preferences, but other notions may be more appropriate for the analysis of collective intentions). We could also determine the dynamic of group aggregation in games: we indeed showed that it was generally in the interest of rational players to become team reasoners, but we did not characterise whether some coalitions are more likely to emerge. In particular, we could investigate whether the grand coalition is likely to form or not, or if several coalitions are likely to survive and enter in competition against each other.
Another possible extension of our results would be to focus on the evolutionary framework, by introducing for instance a noisy observation of the frames of the players. We could also extend our model by considering multilevel selection phenomena, i.e. by assuming that the natural selection operates at the individual level *and* at the level of groups. Although team reasoning may not be evolutionary stable at an individual level if frames are not observed, team reasoners may survive because the groups to which they belong are more fitted.
A last possibility would be to study the impact of public policies on individual preferences, similarly to our analysis in chapter 5. Since public policies impact the structure of the game, and therefore potentially the perception the individuals have of the game, our model can provide a basis for developing models of crowding-out effects: in a situation in which people cooperate because they see the game with a we-frame, the introduction of a monetary incentive may induce those team reasoners to perceive the game with a I-frame, and then to adopt less prosocial behaviours.

Appendix A

# Appendix of chapter 5

## A.1 Lemma 1

We show that:

$$\frac{\partial f_j}{\partial x_i}(x_i|S) = \frac{C_{ij}^{J(S)}}{C_{ii}^{J(S)}}, \tag{A.1}$$

with $f_j(x_i|S)$ the Stackelberg best reply function of player $j$ for $S$, $C_{ij}^{J(S)}$ the $(i;j)$ cofactor of $J(S)$, the Jacobian matrix of the marginal utility functions $U_i^i(x|S)$, evaluated at the Nash equilibrium of $\Gamma_2(S)$.
Consider that all players but $i$ are maximizing their utility functions, i.e. that they play their best reply strategy according to $x_i$; if player $i$ changes her strategy such that $\mathrm{d}x_i \neq 0$, then, we must verify, $\forall j \neq i$ (the different functions are evaluated in $(f_1(x_i); \ldots; f_n(x_i))$, i.e. when all players but $i$ maximize their utility functions):

$$\mathrm{d}U_j^j(x) = 0, \tag{A.2}$$

$$U_j^{ji}\ \mathrm{d}x_i + \sum_{k\neq i} U_j^{jk}\ \mathrm{d}x_k = 0. \tag{A.3}$$

We can rewrite this system of linear equations with $\mathrm{d}x_{-i} = {}^t\{\mathrm{d}x_k\}_{k\neq i}$, and $B_i = {}^t\{u_k^{ki}\,\mathrm{d}x_i\}_{k\neq i}$:

$$J_{ii}\,\mathrm{d}x_{-i} + B_i = 0. \tag{A.4}$$

Since we assumed that $J_{ii}$ is non singular, the system (A.4) has a unique solution, with $J_{ii}^j$ a $(n-1)\times(n-1)$ matrix identical to $J$ except for the column made of $U_k^{kj}$, $\forall k \neq i$ which is replaced by $-B_i$, and without row $i$ and column $i$:

$$\mathrm{d}x_j = \frac{\left|J_{ii}^j\right|}{|J_{ii}|}. \tag{A.5}$$

We can develop the determinant of $J_{ii}^j$ (we suppose that $i < j$ without loss of generality) and add a row and a column at the $i$th place as follows:

$$\left|J_{ii}^j\right| = \begin{vmatrix} U_1^{11} & \dots & U_1^{1,i-1} & 0 & U_1^{1,i+1} & \dots & U_1^{1,j-1} & -U_1^{1,i}\,\mathrm{d}x_i & U_k^{k,j+1} & \dots & U_1^{1n} \\ \dots & \dots & \dots & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ U_{i-1}^{i-1,1} & \dots & U_{i-1}^{i-1,i-1} & 0 & U_{i-1}^{i-1,i+1} & \dots & U_{i-1}^{i-1,j-1} & -U_{i-1}^{i-1,i}\,\mathrm{d}x_i & U_{i-1}^{i-1,j+1} & \dots & U_{i-1}^{i-1,n} \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ U_1^{i+1,1} & \dots & U_{i+1}^{i+1,i-1} & 0 & U_{i+1}^{i+1,i+1} & \dots & U_{i+1}^{i+1,j-1} & -U_{i+1}^{i+1,i}\,\mathrm{d}x_i & U_{i+1}^{i+1,j+1} & \dots & U_{i+1}^{i+1,n} \\ \dots & \dots & \dots & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ U_n^{n1} & \dots & U_n^{n,i-1} & 0 & U_n^{n,i+1} & \dots & U_n^{n,j-1} & -U_n^{ni}\,\mathrm{d}x_i & U_n^{n,j+1} & \dots & U_n^{nn} \end{vmatrix} \tag{A.6}$$

We can then invert the $i^{th}$ with the $j^{th}$ column, and we obtain:

$$\left|J_{ii}^j\right| = (-1)^{i+j}\,|J_{ij}|\,\mathrm{d}x^i. \tag{A.7}$$

We can now rewrite the relation (A.5):

$$\mathrm{d}x^j = \frac{C_{ij}^J}{C_{ii}^J}(f^1(x^i);\dots;f^n(x^i))\,\mathrm{d}x^i. \tag{A.8}$$

This last relation gives us the best reply of player $j$ to a given variation of strategy of player $i$ in order to maximize her utility function when all the other players but $i$ are maximizing their utility functions. We can notice that the primitive of the best reply in terms of variation $\mathrm{d}x_j$ is the Stackelberg best reply function of player $j$, i.e. the strategy $x_j$ which maximizes the utility function $U_j$ for a given strategy of player $i$, knowing the best reply of the other players $k \neq i, j$. We have therefore:

$$\frac{\partial f_j}{\partial x_i}(x_i|S) = \frac{C_{ij}^J}{C_{ii}^J}(f_1(x_i);\dots;f_n(x_i)). \tag{A.9}$$

## A.2 Lemma 2

We now prove that $\frac{\partial f_j}{\partial x_i}$ has the same sign than $U_j^{ji}$ when the two following conditions are verified:

$$(n-1)\left|U_i^{ij}\right| < \left|U_i^{ii}\right|, \qquad \forall j \neq i, \tag{A.10}$$

$$(n-1)\left|U_i^{ij}\right| > \left|U_i^{ik}\right|, \qquad \forall j,k \neq i. \tag{A.11}$$

Condition (A.10) implies a strong form of diagonal dominance for $J$, since it means that the diagonal terms are all significantly greater than all the off-diagonal terms (this condition is identical to diagonal dominance if the off-diagonal terms are identical). Condition (A.11) implies a similar condition, i.e. that the off-diagonal terms are relatively close. In both situations, those conditions mean that there is no player $j$ who has a significantly higher importance for $i$ compared to the other players $k$. We can notice that the condition (A.11) has no sense when $n = 2$, since there does not exist a $k$ different from $i$ and $j$. The condition (A.10) is then sufficient.
We now determine the signs of $C_{ii}^J$ and $C_{ij}^J$. Since $J$ is diagonal dominant, we know that $J_{ii}$ is also diagonal dominant. We also know that, at the Nash equilibrium of $\Gamma_2(S)$, we have $U_i^{ii}(\bar{x}(S)) < 0 \ \forall i \in N$. We have therefore:

$$sign(C_{ii}^J) = (-1)^{n-1}. \tag{A.12}$$

We need now to determine the sign of $C_{ij}^J$. For clarity, we will illustrate our demonstration by focusing on the case $n = 4$. We have the following matrix $J$:

$$J = \begin{pmatrix} U_1^{11} & \dots & U_1^{14} \\ \dots & & \dots \\ U_4^{41} & \dots & U_4^{44} \end{pmatrix} \tag{A.13}$$

Without loss of generality, suppose that $i < j$. We have:

$$C_{ij}^J = (-1)^{(i+j)} \left|J_{ij}\right| \tag{A.14}$$

We now invert lines and columns in $|J_{ij}|$ such that the term $U_j^{ji}$ stand in the first row and first column. This required $(i + j - 3)$ operations: we indeed need $(j-1)$ operations to reach the first column and $(i-2)$ operations to reach the first line (we indeed deleted the $i$th row to obtain $J_{ij}$, and we assumed that $i < j$). $J'_{ij}$ denote the matrix that results from those operations. We obtain:

$$C_{ij}^J = (-1)^{(i+j)}(-1)^{(i+j-3)} \left|J'_{ij}\right|, \tag{A.15}$$

$$C_{ij}^J = -\left|J'_{ij}\right|. \tag{A.16}$$

An interesting property of $J'_{ij}$ is that its first principal minor is necessarily composed by the second order derivatives $U_k^{kk}$, $k \neq i, j$. In the case of $n = 4$, we have for instance:

$$J_{24} = \begin{pmatrix} U_1^{11} & U_1^{12} & U_1^{13} \\ U_3^{31} & U_3^{32} & U_3^{33} \\ U_4^{41} & U_4^{42} & U_4^{43} \end{pmatrix} \tag{A.17}$$

$$J'_{24} = \begin{pmatrix} U_4^{42} & U_4^{41} & U_4^{43} \\ U_1^{12} & U_1^{11} & U_1^{13} \\ U_3^{32} & U_3^{31} & U_3^{33} \end{pmatrix} \tag{A.18}$$

It implies that the first principal minor is row-diagonal dominant. We now operate on the columns of $J'_{ij}$ in order to have $(U_j^{ji}; 0; \ldots; 0)$ as a first row. We obtain:

$$\left|J'_{24}\right| = \begin{vmatrix} U_4^{42} & 0 & 0 \\ U_1^{12} & U_1^{11} - \frac{U_4^{41}}{U_4^{42}} U_1^{12} & U_1^{13} - \frac{U_4^{43}}{U_4^{42}} U_1^{12} \\ U_3^{32} & U_3^{31} - \frac{U_4^{41}}{U_4^{42}} U_3^{32} & U_3^{33} - \frac{U_4^{43}}{U_4^{42}} U_3^{32} \end{vmatrix} \tag{A.19}$$

We can now check that, under the assumptions (A.10) and (A.11), the first principal minor is still diagonal dominant, and the diagonal terms are still negative. We have then:

$$sign(\left|J'_{ij}\right|) = sign\left(U_j^{ji}(-1)^{n-2}\right). \tag{A.20}$$

We obtain:

$$sign(C_{ij}^J) = -sign(\left|J'_{ij}\right|), \tag{A.21}$$

$$sign(C_{ij}^J) = (-1)^{n-1} sign(U_j^{ji}). \tag{A.22}$$

We can now complete our proof and determine the sign of $\frac{\partial f_j}{\partial x_i} = \frac{C_{ij}^J}{C_{ii}^J}$:

$$sign\left(\frac{\partial f_j}{\partial x_i}\right) = \frac{sign\left(C_{ij}^J\right)}{sign\left(C_{ii}^J\right)}, \tag{A.23}$$

$$sign\left(\frac{\partial f_j}{\partial x_i}\right) = sign(U_j^{ji}). \tag{A.24}$$

## A.3 Lemma 3

We now show that, under the same assumptions than lemma 2, we have

$$\sum_{j\neq i}\left|C_{ij}^{J(S)}\right| < \left|C_{ii}^{J(S)}\right|, \tag{A.25}$$

$$\iff \sum_{j\neq i}\left|\frac{\partial f_j}{\partial x_i}\right| < 1. \tag{A.26}$$

We know that:

$$C_{ii}^J = \sum_{k\neq i} U_k^{kj} C_{kj}^{J_{ii}}, \qquad \forall j \neq i, \tag{A.27}$$

$$C_{ii}^J = \sum_{j\neq i}\left[\frac{1}{n-1}\sum_{k\neq i} U_k^{kj} C_{kj}^{J_{ii}}\right], \tag{A.28}$$

$$C_{ij}^J = -\sum_{k\neq i} U_k^{ki} C_{kj}^{J_{ii}} \tag{A.29}$$

Without loss of generality, suppose that $C_{ii}^J$ is positive (which is true if $n$ is odd). We therefore have:

$$\left|C_{ii}^J\right| = C_{ii}^J, \tag{A.30}$$

$$\left|C_{ij}^J\right| = \sum_{k\neq i} U_k^{ki} C_{kj}^{J_{ii}}. \tag{A.31}$$

We therefore obtain:

$$\left|C_{ii}^J\right| - \sum_{j\neq i}\left|C_{ij}^J\right| = \sum_{j\neq i}\sum_{k\neq i}\left(\frac{U_k^{kj}}{n-1} - U_k^{ki}\right) C_{kj}^{J_{ii}}. \tag{A.32}$$

(A.25) is true if and only if:

$$\sum_{j\neq i}\sum_{k\neq i}(\frac{U_k^{kj}}{n-1}-U_k^{ki})C_{kj}^{J_{ii}}>0. \tag{A.33}$$

If we multiply by $(n-1)$ and divide on both sides by $C_{jj}^{J_{ii}}$ (negative by construction, since we assumed $C_{jj}^{J}>0$), we obtain:

$$\sum_{j\neq i}\left[U_j^{jj}-(n-1)U_j^{ji}+\sum_{k\neq i,j}(U_k^{kj}-(n-1)U_k^{ki})\frac{C_{kj}^{J_{ii}}}{C_{jj}^{J_{ii}}}\right]<0. \tag{A.34}$$

We can now check that under conditions (i) and (ii), this condition is verified (the second term is well negative, since $\frac{C_{kj}^{J_{ii}}}{C_{jj}^{J_{ii}}}$ has the same sign than $U_j^{jk}$ by lemma 2). (A.25) is therefore true.

## A.4 Proposition 1

We show that $(\bar{x};I_n)$ is a SPEC of $\Gamma$ if and only if:

(i) either $\forall i,j\in N$, $\Pi_i^j(\bar{x})\Pi_j^i(\bar{x})=0$,

(ii) or $\forall i,j\in N$, $i\neq j$, $\Psi_i^i(\bar{x})=0$,

with $\bar{x}\in X$ the Nash equilibrium of $\Gamma$, and $I_n$ a matrix in $\mathbb{R}^{n\times n}$ such that $\sigma_{ij}\neq 0$ if and only if $i=j$.
$\forall S\in\mathbb{R}^{n\times n}$, there exists a unique Nash equilibrium for $\Gamma_2(S)$ $\bar{x}\in X$ that verifies, $\forall i\in N$:

$$U_i^i(\bar{x}|S)=0, \tag{A.35}$$

$$\sum_{j\in N}\sigma_{ij}\Pi_j^i(\bar{x})=0. \tag{A.36}$$

By definition of the Stackelberg best reply function, the indirect payoff function $V_i:S\mapsto\mathbb{R}$ can be rewritten as follows:

$$V_i(S) = \Pi_i(\bar{x}(S)), \tag{A.37}$$

$$V_i(S) = \Pi_i(f_1(\bar{x}_i(S)); \ldots; \bar{x}_i(S); \ldots; f_n(\bar{x}_i(S))), \tag{A.38}$$

$$V_i(S) = \Psi_i(\bar{x}_i|S). \tag{A.39}$$

The relation (A.39) implies that, at the Nash equilibrium of $\Gamma_1$, player $i$ maximises her Stackelberg payoff function when she maximises her indirect payoff $V_i$. We must therefore verify, at the Nash equilibrium of $\Gamma_1$:

$$\frac{\partial V_i}{\partial \sigma_{ij}}(\bar{S}) = \Psi_i^i(\bar{x}_i|\bar{S})\frac{\partial \bar{x}_i}{\partial \sigma_{ij}} \;=\; 0, \qquad \forall j \in N, \tag{A.40}$$

$$\text{i.e. either} \qquad \frac{\partial \bar{x}_i}{\partial \sigma_{ij}}(\bar{S}) = 0, \qquad \forall j \in N, \tag{A.41}$$

$$\text{or} \qquad \Psi_i^i(\bar{x}_i|\bar{S}) = 0.. \tag{A.42}$$

If (A.41) is not true, then $I_n$ is a first stage game equilibrium if and only if (A.42) holds for $\bar{x}$, the Nash equilibrium of $\Gamma$. We have therefore proven the condition (ii) of proposition 1.

We now determine the conditions under which the conditions (A.41) holds. We must therefore characterise $\bar{x}(S)$, the Nash equilibrium of $\Gamma_2(S)$. We identify the best reply of player $i$, $\forall i \neq j$, when a player $j$ unilaterally changes her strategy $S_j$. We consider here the differential of $U_i^i$, and look for the reactions $\mathrm{d}x_i$ that verify $\mathrm{d}U_i^i(\bar{x}) = 0$, $\forall i \in N$. We have the following relations:

$$\mathrm{d}U_i^i(\bar{x}) = 0, \qquad \forall i \in N, \tag{A.43}$$

$$\sum_{j \in N} \left[ U_i^{ij}(\bar{x})\ \mathrm{d}x_j + \Pi_j^i(\bar{x})\ \mathrm{d}\sigma_{ij} \right] = 0, \qquad \forall i \in N. \tag{A.44}$$

We solve this system of linear equations in $\mathrm{d}x_i$:

$$\begin{pmatrix} U_1^{11}(\bar{x}) & \ldots & U_1^{1n}(\bar{x}) \\ \ldots & & \ldots \\ U_n^{n1}(\bar{x}) & \ldots & U_n^{nn}(\bar{x}) \end{pmatrix} \begin{pmatrix} \mathrm{d}x_1 \\ \ldots \\ \mathrm{d}x_n \end{pmatrix} + \begin{pmatrix} \sum_{j \in N} \Pi_j^1(\bar{x})\,\mathrm{d}\sigma_{1j} \\ \ldots \\ \sum_{j \in N} \Pi_j^n(\bar{x})\,\mathrm{d}\sigma_{nj} \end{pmatrix} = 0, \tag{A.45}$$

$$J(S)\,\mathrm{d}x + \mathrm{d}A = 0, \tag{A.46}$$

with $\mathrm{d}x = {}^t\{\mathrm{d}x_i\}_{i\in N}$ the column vector of strategies' variations; $\mathrm{d}A = {}^t\{\mathrm{d}A_i\}_{i\in N}$. We make the additional assumption that $J(S)$ and its minors $J_{ii}(S)$ are generically non singular $\forall S \in \mathbb{R}^{n\times n}$. The system (A.46) has therefore a unique solution (for notational convenience, we do not mention on which set of parameters $S$ $J$ is defined, unless a confusion is possible):

$$\mathrm{d}x_i = \frac{|J^i|}{|J|} \qquad \forall i \in N, \tag{A.47}$$

with $J^i$ a $n \times n$ matrix identical to $J$, except for the $i^{th}$ column which is replaced by $-\mathrm{d}A$. We deduce the following relations:

$$\mathrm{d}x_i = -\frac{\sum_{k\in N} C^J_{ki}\,\mathrm{d}A^k}{|J|}, \tag{A.48}$$

$$\Longrightarrow \qquad \frac{\partial \bar{x}_i}{\partial \sigma_{ik}}(S) = -\,\Pi^i_k \frac{C^J_{ii}}{|J|}(\bar{x}(S)) \qquad \forall S \in \mathbb{R}^{n\times n}. \tag{A.49}$$

We can therefore see the condition (A.41) implies:

$$\Pi^i_k(\bar{x}) = 0 \qquad \forall k \in N. \tag{A.50}$$

This last condition means that the strategy profile that maximizes the utility function $U_i$ of player $i$ also maximizes her own payoff $\Pi_i$ as well as the payoff of all the other players $j \neq i$ (or minimizes it). It means therefore that, if $\forall i, k \in N$, $\Pi^i_k(\bar{x}) = 0$ at Nash equilibrium, then $(\bar{x}; I_n)$ is a SPEC.

## A.5 Proposition 2

We prove here that $\bar{S}$ is a Nash equilibrium of the first stage game $\Gamma$ when:

$$\bar{\sigma}_{ij} = \frac{\Pi^j_i}{\Pi^i_j}(\bar{x})\frac{\partial f_j}{\partial x_i}(\bar{x}_i|\bar{S}). \tag{A.51}$$

We look for conditions under which (A.42) is verified at the first stage game equilibrium. We can therefore rewrite the first order condition of the first stage game equilibrium (A.40):

$$\sum_{j\in N} \Pi_i^j \frac{\partial f_j}{\partial x_i}(\bar{x}_i(S)) = 0. \tag{A.52}$$

Combining equations (A.36) and (A.52), we can obtain an expression of the parameters $\sigma_{ij}$ at equilibrium:

$$\sum_{j\in N} \Pi_i^j \frac{\partial f_j}{\partial x_i} = \sum_{j\in N} \sigma_{ij} \Pi_j^i. \tag{A.53}$$

We can then suggest the following specification for the first stage game equilibrium:

$$\sigma_{ij} = \frac{\Pi_i^j}{\Pi_j^i} \frac{\partial f_j}{\partial x_i}. \tag{A.54}$$

Note that, since we are maximising the Stackelberg function in the first stage game $\Gamma_1$, we only have $n$ equations to determine the $n^2$ parameters $\sigma_{ij}$. Although other specifications were possible, we chose here to define $\sigma_{ij}$ as a function of the Stackelberg best reply of player $j$ when $i$ is the leader, since it captures the idea that the behaviour of $i$ towards $j$ fundamentally depends on the way $j$ reacts when $i$ changes her strategy.

## A.6 Proposition 3

We prove that, under the following assumptions:

(i) $\left|U_i^{ii}\right| > (n-1)\left|U_i^{ij}\right|$,

(ii) $\left|U_i^{ik}\right| < (n-1)\left|U_i^{ij}\right|$,

(iii) $\left|\Pi_j^{ji}\right| \geq \left|\Pi_k^{ji}\right|$,

a symmetric SPEC $(\bar{x}; \bar{S})$ verifies:

$$sign\left(\bar{\sigma}_{ij}\right) = sign\left(\Pi_j^{ji}(\bar{x}(S))\right), \qquad \forall j \neq i. \tag{A.55}$$

A symmetric SPEC implies that $\Pi_i^j(\bar{x}(\bar{S})) = \Pi_j^i(\bar{x}(\bar{S}))$. The optimal weights $\bar{\sigma}_{ij}$ are therefore:

$$\bar{\sigma}_{ij} = \frac{\partial f_j}{\partial x_i}(\bar{x}(\bar{S})). \tag{A.56}$$

Lemma 3 ensures that, under conditions (i) and (ii), we have:

$$\sum_{j \neq i} |\bar{\sigma}_{ij}| < 1. \tag{A.57}$$

We can then easily deduce the following relation, when condition (iii) is verified:

$$\left|\Pi_i^{ij}\right| > \left|\sum_{k \neq i} \sigma_{ij} \Pi_k^{ij}\right|, \tag{A.58}$$

$$sign(U_i^{ij}) = sign(\Pi_i^{ij}). \tag{A.59}$$

By lemma 2, we know that $\bar{\sigma}_{ij}$ has the same sign than $U_j^{ji}$. We have therefore, at a symmetric SPEC:

$$sign\left(\bar{\sigma}_{ij}\right) = sign\left(\Pi_j^{ji}(\bar{x}(S))\right), \qquad \forall j \neq i. \tag{A.60}$$

Appendix B

# Appendix of chapter 6

## B.1 Proposition 4

Since a team $S$ chooses a strategy profile $x_{S,S}$ rather than each component $x_{i,S}$, it is possible to avoid issues of coordination within the team when maximising the function $U_S$. For a given incomplete strategy profile $\bar{x}_{-(S;S)}$ the coalition $S$ must choose $x_{S,S}$ such that it maximises $U_S$: the supervisor then decides the optimal profile for the team, and — since all the players in $S$ have the same preferences — maximising $U_S$ implies that no sub-coalition $T \subseteq S$ can unilaterally deviate from $\bar{x}_{S,S}$ to obtain a higher payoff. The two notions of UTI-equilibrium for $\Gamma$ and team-proof Bayesian Nash equilibrium for $BG(\Gamma)$ are therefore strategically equivalent.

## B.2 Proposition 5

The proof is similar to the one of proposition 4: For a given incomplete strategy profile $\bar{x}_{-(S;S)}$, the supervisor of $S$ in the UTI must choose a collective profile $\bar{x}_{S,S}$ that collectively maximises $U_S$. Since all the players in $S$ have the same preferences, the profile that maximises $U_S$ gives the highest possible payoff to all the players in the game $G(\Gamma)\backslash\bar{x}_{-(S;S)}$, and is therefore a strong Nash equilibrium of $G(\Gamma)\backslash\bar{x}_{-(S;S)}$.

## B.3 Proposition 6

The proof follows the proof of proposition 2: several agents (here coalitions) are indeed choosing their optimal weights given their objective functions $\bar{U}_S$. The only difference is now that the strategy space of a coalition is not a subset of $\mathbb{R}$ but of $\mathbb{R}^{|S|}$.

At the second stage equilibrium, the supervisor of $S$ is maximising the objective function $U_S$:

$$\frac{\partial \bar{U}_S}{\partial x_i} + \sum_{j \in M \setminus S} \sigma_{Sj} \frac{\partial \bar{U}_j}{\partial x_i} = 0, \qquad \forall i \in S. \tag{B.1}$$

In the first stage, the supervisor must choose the weights $\sigma_{Sj}$ such that the resulting equilibrium maximises the Stackelberg payoff of the team, i.e. the function $\bar{U}_S$ taking into account the Stackelberg best reply of the outsiders:

$$\frac{\partial \bar{U}_S}{\partial x_i} + \sum_{j \in M \setminus S} \frac{\partial f_j}{\partial x_i} \frac{\partial \bar{U}_i}{\partial x_j} = 0, \qquad \forall i \in S, \tag{B.2}$$

with $f_j(x_i)$ the Stackelberg best reply of outsiders for a strategy $i \in S$, and given strategies of $j \in S$.
Note that $\frac{\partial f_j}{\partial x_i}$ has not exactly the same expression than the one given by lemma 1, since the players $j \in S \setminus i$ are not playing their best reply. By applying the same proof than in lemma 1, we obtain:

$$\frac{\partial f_j}{\partial x_i} = \frac{C_{ij}^{J(S)}}{C_{ii}^{J(S)}}, \tag{B.3}$$

with $J(S)$ a $(m - |S| + 1) \times (m - |S| + 1)$ matrix identical to $J(G(\Gamma))$ in which the lines and rows corresponding to the players $j \in S \setminus i$ have been deleted.

Combining conditions (B.1) and (B.2), we can then deduce an expression of the optimal weights $\sigma_{Sj}$. As in proposition 2, we have less equilibrium conditions than parameters to determine ($|S|$ conditions only to determine $(n2^{n-1} - |S|)$ parameters). We are therefore looking for parameters $\sigma_S$ such that:

$$\sum_{j \in M \setminus S} \sigma_{Sj} \frac{\partial \bar{U}_j}{\partial x_i} = \sum_{j \in M \setminus S} \frac{\partial f_j}{\partial x_i} \frac{\partial \bar{U}_i}{\partial x_j}. \tag{B.4}$$

However, we cannot offer here a general and clear solution as the one suggested in proposition 2, by associating to each player his individual effect on the resulting equilibrium. This is however only an issue of readability, and the core result that coalitions tend to put positive weights in supermodular games will still be true. So as to offer a presentable form of optimal weights, we restrict ourselves to aggregative games, i.e. games such that $\Pi_i(x) = \Pi_i(x_i; \sum_{j \neq i} x_j)$. This last condition indeed implies that a player $j \notin S$ reacts in the exact same way whoever the deviant player

is in $S$ (the deviation indeed only impact the total aggregate of strategies of $j$'s outsiders, and not the strategy of her own coalition):

$$\begin{cases} \dfrac{\partial f_j}{\partial x_i} = \dfrac{\partial f_j}{\partial x_k}, & \forall i,k \in S, j \notin S, \\ \dfrac{\partial \bar{U}_j}{\partial x_i} = \dfrac{\partial \bar{U}_j}{\partial x_k}, & \forall i,k \in S, j \notin S. \end{cases} \tag{B.5}$$

We can then suggest the following weights as a solution to (B.3):

$$\bar{\sigma}_{S,j} = \frac{\frac{\partial \bar{U}_S}{\partial x_j}}{\frac{\partial \bar{U}_j}{\partial x_i}} \frac{C_{ij}^{J(S)}}{C_{ii}^{J(S)}}. \tag{B.6}$$

## B.4 Proposition 7

*Proof*: we study the conditions under which $i$ has an interest in team reasoning with a coalition $S \subseteq N$, knowing that all the other players are I-reasoners. Suppose that all the players but $i$ have a I-frame, and are therefore maximising their expected utility. If $i$ chooses to adopt a frame $S_i \neq i$, then she will play a UTI in which she is the only team reasoner (in particular, the other players $j \in S_i$ do not identify with $S_i$). $i$ must therefore choose the collective preferences of $S_i$ such that the joint product of the individual benefice from team reasoning is maximised among the team reasoners of $S_i$. Since $i$ is the only team reasoner, she therefore chooses collective preferences such that they maximise *in fine* her own utility.
$i$ therefore chooses collective preferences such that the resulting UTI-equilibrium $\tilde{x}$ maximises $\Pi_i(x)$, knowing that the other players are playing their best reply. $i$ has therefore an incentive to choose a frame $S_i \neq i$ if and only if she can choose preferences such that the resulting equilibrium maximises her Stackelberg payoff.
Therefore, $\{\bar{S}_i = i\}_{i \in N}$ is a Nash equilibrium of $G_0 = \langle N, \mathcal{S}, V \rangle$ if and only if, $\forall i \in N$:

$$\Psi_i(\bar{x}_i) \geq \Psi_i(x_i), \qquad \forall x_i \in X_i. \tag{B.7}$$

If (B.4) is not verified, then it means that there exists a player $i$ such that, when player $i$ is a Stackelberg leader, she obtains a strictly higher payoff than at Nash equilibrium: $i$ would therefore have an incentive in becoming a team reasoner so as to choose collective preferences that implements *in fine* her Stackelberg equilibrium.

## B.5 Proposition 8

*Proof*: (i) we firstly prove that $\{\bar{S}_i = N\}_{i \in N}$ is always a Nash equilibrium of $G_0$. We must therefore show that, when both players are team reasoning, no player can benefit from being the Stackelberg follower of the other (i.e. of letting the other player team reasoning alone). Let $\tilde{x}^i$ denote the Stackelberg equilibrium when $i$ is the leader. Suppose firstly that $i$ gets less than at Nash equilibrium when she is the follower of $j$ (as in a submodular game). In this situation, since the equilibrium when both players team reason gives at least their Nash payoff to both players, then $i$ is better off by choosing to team teal reason when $j$ team reasons. Suppose now that $i$ gets more at $\tilde{x}^j$ than at $\bar{x}$: if the supervisor plays $\tilde{x}^j$ as the equilibrium for the team when both players team reason, then no player has an incentive in becoming a I-reasoner when the other team reasons: $i$ is indeed indifferent between team reasoning and not (because it leads her to the same outcome), while $j$'s payoff is maximised (she indeed achieved her Stackelberg payoff). $\{\bar{S}_i = N\}_{i \in N}$ is therefore necessarily a Nash equilibrium of $G_0$.

(ii) we now prove that only a Nash equilibrium can be played as a subgame perfect UTI-equilibrium if and only if no one has a first mover advantage, and if the equilibrium is Pareto optimal with no Pareto equivalent. We know by proposition 7 that $\{i; i\}$ is a Nash equilibrium if and only no player can benefit from a first mover advantage. We are now looking for a condition of uniqueness of this equilibrium (a condition under which no profile can be played as the result of an optimal strategic commitment in the first stage game). We can then simply notice that, if $\bar{x}$ is not Pareto optimal, then the grand coalition would be better off by deviating to the Pareto superior profile. Furthermore, if $\bar{x}$ is Pareto optimal, then it is immune against a collective deviation: no player, either as a Stackelberg leader or as member of the grand coalition can therefore achieve more than at Nash equilibrium (the absence of Pareto-equivalence outcomes ensures that the grand coalition cannot deviate to a profile that pays the same than $\bar{x}$). It is therefore not possible to have a subgame perfect UTI-equilibrium such that the resulting strategy profile is not the Nash equilibrium $\bar{x}$.

Appendix C

# Appendix of chapter 7

## C.1 Proposition 9

*Proof*: we can firstly notice that, if $p_{DD} \neq 0$, the share of CC-types necessarily decreases (since they are strictly outperformed by DD and CD-types): $\dot{p}_{CC} < 0$. The share of CD-types increases over time if and only if:

$$\Pi^t_{CD} > \bar{\Pi}^t, \tag{C.1}$$

$$\Leftrightarrow \quad p^t_{CC}(T+S-2P) < (1-p^t_{DD})(R-P), \tag{C.2}$$

$$\Leftrightarrow \quad \frac{p^t_{CD}}{p^t_{CC}} > \frac{T+S-R-P}{R-P}. \tag{C.3}$$

Condition (C.3) implies that, if the share of unilateral cooperators is sufficiently low relatively to the share of conditional cooperators (or alternatively, if $T+S < P+R$), then the share of conditional cooperators will progressively increase. We can now notice that $p^t_{CD}$ necessarily grows faster than $p^t_{CC}$:

$$\dot{\left(\frac{p^t_{CD}}{p^t_{CC}}\right)} = \frac{\dot{p}_{CD}p^t_{CC} - \dot{p}_{CC}p^t_{CD}}{(p^t_{CC})^2}, \tag{C.4}$$

$$\Leftrightarrow \quad \dot{\left(\frac{p^t_{CD}}{p^t_{CC}}\right)} = \frac{p^t_{CD}}{p^t_{CC}}\frac{p^t_{DD}(P-S)}{\bar{\Pi}^t} > 0. \tag{C.5}$$

Condition (C.5) implies that, if there exists a date $t^*$ such that (C.3) is verified, then (C.3) will be verified $\forall t \geq t^*$ (the ratio $\frac{p^t_{CD}}{p^t_{CC}}$ will indeed continue to strictly increase over time). The system will then asymptotically converge to $(p_{CC}=0, p_{DD}=0, p_{CD}=1)$.

Suppose on the contrary that (C.3) is not verified. The share of DD-types then necessarily increases (since the shares of CC and CD-types decrease), i.e. $\dot{p}_{DD} > 0$. We have therefore:

$$\dot{p}_{DD}(P-S) > 0, \tag{C.6}$$

$$\dot{\Pi}^t < 0. \tag{C.7}$$

It then follows from (C.5), (C.6) and (C.7) that $\left(\frac{\dot{p^t_{CD}}}{p^t_{CC}}\right)$ increases with $t$, i.e. that the ratio $\frac{p^t_{CD}}{p^t_{CC}}$ is convex in $t$ when (C.3) is not verified. Since the ratio $\frac{p^t_{CD}}{p^t_{CC}}$ is increasing and convex, we are certain that it will pass any finite threshold. There therefore necessarily exists a date $t^*$ such that, $\forall t \geq t^*$, the condition (C.3) is verified.
It is not certain that the trajectory of $p^t_{CD}$ is monotonic: while $p^t_{CC}$ necessarily decreases over time, $p^t_{CD}$ also decreases $\forall t < t^*$. It is necessary to wait for the quasi-extinction of CC-types for CD-types to be able to outperform DD-types. $\forall t > t^*$, $p^t_{CC}$ and $p^t_{DD}$ decreases while $p^t_{CD}$ strictly increases. The system (7.5) asymptotically reach $(p_{CC} = 0, p_{DD} = 0, p_{CD} = 1)$.

## C.2 Proposition 10

*Proof.* Suppose by way of contradiction that there exists a set of reply functions $R$ such that there is no $s$-cycle. It means that, whatever the initial strategy profile $s \in S$ is, it is not possible to have a sequence of best reply that leads us to this same initial state. Since there exists a finite number of strategy profiles, it is not possible to build an infinite sequence $C$ such that all the strategy profiles are distinct. $\forall R \in \mathcal{R}$, there therefore necessarily exists a $s$-cycle.

## C.3 Proposition 11

*Proof.* (i) We adapt the notions of Stackelberg reply function to our framework with heuristics:

**Definition 15.** *The function $f_j : S_i \mapsto S_j$ is the Stackelberg reply function of player $j$ if and only if, $\forall s_i \in S_i$:*

$$f_j(s_i) = R_j(f_1(s_i); \ldots; s_i; \ldots; f_n(s_n)), \tag{C.8}$$

*with $f_k : S_i \mapsto S_k$ the Stackelberg best reply function of player $k$, $\forall k \neq i, j$.*

For notational convenience, we will assume that there exists a unique fixed point for any multivalued mapping $R \in \mathcal{R}$, ensuring that the functions $f_j$ are well defined on a non empty subset of $S_i$, $\forall i, j \in N$. We should otherwise define the image of the Stackelberg reply function as the set of strategies $s_j \in C_j(R_{-i})$, i.e. the set of possible replies of player $j$ when $s_i$ is fixed.

(i) We firstly look for the conditions under which PMB is as-if evolutionary stable, i.e. stable against the invasion of a single population. $\theta^{PMB} \in \Delta(\mathcal{R})$ is neutrally stable if and only if, $\forall i \in N$:

$$\int V_i(BR_i; R_{-i}) \, \mathrm{d}\theta_{-i}^{PMB} \geq \int V_i(R_i; R_{-i}) \, \mathrm{d}\theta_{-i}^{PMB}, \qquad \forall R_i \in \mathcal{R}_i, \tag{C.9}$$

$$\iff \qquad \Pi_i(\bar{s}) \geq \Pi_i(\tilde{s}), \tag{C.10}$$

with $\bar{s} \in C(BR)$ and $\forall \tilde{s} \in C(R_i; BR_{-i})$. Since $\bar{s}$ is a fixed point of $BR$, we have $\forall i \in N$:

$$f_j(\bar{s}_i) = \bar{s}_j, \qquad \forall j \in N. \tag{C.11}$$

We can therefore rewrite the condition (C.10) as follows, $\forall i \in N$, $\forall R_i \in \mathcal{R}_i$:

$$\Pi_i(f_1(\bar{s}_i); \ldots; f_n(\bar{s}_i)) \geq \Pi_i(f_1(\tilde{s}_i); \ldots; f_n(\tilde{s}_i)), \tag{C.12}$$

$$\iff \qquad \Psi_i(\bar{s}_i) \geq \Psi_i(\tilde{s}_i). \tag{C.13}$$

The condition (C.13) means that "choosing" the reply function $R_i$ that maximises one's fitness implies choosing the equilibrium $\bar{s}$ that maximises one's Stackelberg function, knowing the reply functions of the other players. A direct corollary of this condition is that, if a player can improve her material payoff by becoming a Stackelberg leader — knowing that the other players are maximising their material payoff — then $\theta^{PMB}$ cannot be neutrally stable: a mutant whose reply function put her in a position of Stackelberg leader will indeed outperform payoff maximisers within her population. Furthermore, if there exists a Nash equilibrium which is also a Stackelberg equilibrium when $i$ is the leader, $\forall i \in N$,

then no player can benefit from changing unilaterally her heuristic when they play this equilibrium. $\theta^{PMB}$ is neutrally stable if and only if no one can benefit from a first mover advantage, i.e. if and only if $\nexists \tilde{s} \in St(i)$ such that $\Pi_i(\tilde{s}) \geq \Pi_i(\bar{s})$.
Notice however that our definition of "as-if evolutionary stability" excludes neutral mutants when the resulting equilibrium is not the initial Nash equilibrium (the only possible neutral mutants therefore behave as if they were maximising their payoff). So as to ensure that $\theta^{PMB}$ is as-if evolutionary stable, we must also assume that there does not exist a Stackelberg equilibrium $\tilde{s} \in St(i)$, $\tilde{s} \neq \bar{s}$ such that $\Pi_i(\tilde{s}) = \Pi_i(\bar{s})$, i.e. that $\bar{s}$ is the only Stackelberg equilibrium for $i$, $\forall i \in N$.

(ii) We now investigate the conditions under which $\theta^{PMB}$ is robust against indirect invasions. We can firstly notice that if $\bar{s} \in C(BR)$ does not belong to the $\phi$-core of $\Gamma$, then there exists a coalition $M \subset N$ such that all the players of $M$ can obtain a higher payoff (compared to $\bar{s}$) when $M$ is a Stackelberg leader and the players $N \setminus M$ act as singletons. Consider the following reply function for each player $i \in M$, with $\tilde{s} \in St(M)$, and $\Pi_M(\tilde{s}) \geq \Pi_M(\bar{s})$:

- $\tilde{R}_i(s_{-i}) = BR_i(s_{-i})$, if $\exists j \in M$ s.t. $s_j \neq \tilde{s}_j$,
- $\tilde{R}_i(s_{-i}) = \tilde{s}_i$, if $\forall j \in M$, $s_j = \tilde{s}_j$.

$\tilde{R}_i$ can be described as following: "maximise your material payoff in every circumstances, unless all of your teammates play the optimal strategy of the coalition $M$". We can easily notice that $\tilde{R}_i$ weakly dominates $BR_i$: if the other members of the coalition $M$ do not make their part of the optimal collective profile, $i$ maximises her material payoff; all the other players in $M$ therefore knows that there is no cooperation within the coalition and also maximise their material payoff; they therefore play the Nash equilibrium and obtain the same payoff than if their reply functions were $BR_j$, $\forall j \in M$. However, if all the other players in $M$ play the optimal strategy for the coalition, then $i$ also plays the best strategy for the coalition. Since $M$ is a Stackelberg leader, the material payoff of its members is higher than at Nash equilibrium: on average, $i$ obtains a higher payoff than if her reply function were $BR_i$. Although the entry of a mutant $\tilde{R}$ within the population $i$ will not have an effect on the resulting equilibrium, it gives the opportunity to the entry of other mutants $\tilde{R}$ in the populations $j \in M$: if the different populations $j \in M$ are invaded (either simultaneously or sequentially) by mutants $\tilde{R}$, then those players will outperform payoff maximisers. This therefore means that if there exists a coalition who can benefit from becoming a Stackelberg leader, then is not

robust against indirect invasion.

We now consider the case in which $\bar{s}$ belongs to the $\phi$-core but not to the strict $\phi$-core: there therefore exists a coalition $M \subset N$ that can reach a Stackelberg equilibrium offering exactly to all the members of $M$ their Nash payoff. By following a similar heuristics than $\tilde{R}$, it is therefore possible that mutants from this type survive, since they achieve the same payoff than payoff maximisers, although the resulting equilibrium is not $\bar{s}$. In this situation, the postentry distribution is only neutrally stable, and not as-if evolutionary stable. $\bar{s}$ is therefore not RAII.

The last step of our demonstration consists in showing that if $\bar{s} \in C(BR)$ is in the strict $\phi$-core of $\Gamma$, then $\theta^{PMB}$ is RAII. We can simply notice that, if a group of mutants enter several populations, then they will not be outperformed by payoff maximisers if and only if they get at least the Nash payoff. Since $\bar{s}$ is in the strict $\phi$-core, we know by definition that there is no Stackelberg equilibrium different from $\bar{s}$ such that the players of a coalition $M \subset N$ can achieve at least their Nash payoff. Mutants will therefore not be outperformed if and only if they play the Nash equilibrium, i.e. they behave as if they were maximising their payoff.

## C.4 Proposition 12

*Proof*: we consider two vectors of types $\bar{R}$ and $\hat{R}$ such that: (i) $\bar{\theta}$ is as-if evolutionary stable, (ii) $\hat{R}_i$ induces the exact same choice than $\bar{R}_i$, unless all the players have the type $\hat{R}$. Since $supp(\hat{\theta})$ is a singleton, we do not have to check whether all the profiles support of $\hat{\theta}$ have the same average payoff. Consider now the invasion of population $j$ by individuals of type $R_j \neq \hat{R}_j$: the players are then computing their strategy as if they were of type $\bar{R}_i$. Since $\bar{R}$ is as-if evolutionary stable, it implies that no mutant can successfully invade the population $j$, $\forall j \in N$. $\hat{\theta}$ is therefore as-if evolutionary stable.

# Bibliography


Akerlof, G. (1970). "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism". *Quaterly Journal of Economics*, 84:488–500.

Alchian, A. (1950). "Uncertainty, Evolution and Economic Theory". *Journal of Political Economy*, 58:211–221.

Andersen, S. and Ross, L. (1984). "Self-Knowledge and Social Inference: I. The Impact of Cognitive/Affective and Behavioral Data". *Journal of Personality and Social Psychology*, 46:280–293.

Arena, R. and Gloria-Palermo, S. (2001). "Evolutionary Themes in the Austrian Tradition: Menger, Wieser and Schumpeter on Institutions and Rationality". In Garrouste, P. and Ioannides, S., editors, *Evolution and Path Dependence in Economic Ideas: Past, Present*. Aldershot: Edward Elgar.

Aristotle (350BC). *Nicomachean Ethics*. Batoche Books Kitchener. Translated by W.D. Ross (1999).

Arrow, K. (1951). *Social Choice and Individual Values*. Yale University Press.

Arrow, K. (1995). "A Note on Freedom and Flexibility". In Basu, K., Pattanaik, P., and Suzumura, K., editors, *Choice, Welfare and Development: A Festschrift in Honour of Amartya K. Sen*, pages 7–16. Oxford: Oxford University Press.

Asch, S. (1955). "Opinions and Social Pressure". *Scientific American*, 193(5):31–35.

Aumann, R. J. (1987). Correlated Equilibrium as an Expression of Bayesian Rationality. *Econometrica*, 55(1):1–18.

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.

Bacharach, M. (1991). "Games with Concept-sensitive Strategy Spaces". International Conference on Game Theory, Florence.

Bacharach, M. (1993). "Variable Universe Games". In Binmore, K., Kirman, A., and Tani, P., editors, *Frontiers of Game Theory*. MIT Press.

Bacharach, M. (1995). "Co-Operating without Communicating". Working paper, Institute of Economics and Statistics, University of Oxford.

Bacharach, M. (1997). " 'We' equilibria: a Variable Frame Theory of Cooperation". Working paper, Institute of Economics and Statistics, University of Oxford.

Bacharach, M. (1999). "Interactive Team Reasoning: a Contribution to the Theory of Co-operation". *Research in Economics*, 53:117–147.

Bacharach, M. (2003). "Framing and Cognition: The Bad News and the Good". In Dimitri, N., Basili, M., and Gilboa, I., editors, *Cognitive Processes and Economic Behaviour*. London: Routledge.

Bacharach, M. (2006). *Beyond Individual Choice. Teams and Frames in Game Theory*. Princeton University Press. Edited by Natalie Gold and Robert Sugden.

Bacharach, M. and Bernasconi, M. (1997). "The Variable Frame Theory of Focal Points: An Experimental Study". *Games and Economic Behavior*, 19:1–45.

Bacharach, M. and Hurley, S. (1991). "Issues and Advances in the Foundations of Decision Theory". In Bacharach, M. and Hurley, S., editors, *Foundations of Decision Theory*. Oxford: Blackwell Publisher.

Bardsley, N. (2007). "On Collective Intentions: Collective Action in Economics and Philosophy". *Synthese*, 157(2):141–159.

Bardsley, N., Mehta, J., Starmer, C., and Sugden, R. (2006). "Explaining Focal Points: Cognitive Hierarchy Thery versus Team Reasoning". *Economic Journal*, 120:40–79.

Barrett, S. (1999). "A Theory of Full International Cooperation". *Journal of Theoretical Politics*, 11(4):519–541.

Barrett, S. (2003). "Environment and Statecraft: The Strategy of Environmental Treaty-Making". *American Economic Review*, 90:166–193.

Barrett, S. (2007). *Why Cooperate? The Incentive to Supply Global Public Goods*. Oxford University Press.

Battalio, R., Samuelson, L., and Von Huyck, J. (2001). "Optimization Incentives and Coordination Failure in Laboratory Stag Hunt Games". *Econometrica*, 63(3):749–764.

Baujard, A. and Gilardone, M. (2015). "Sen is not a Capability Theorist". Available at SSRN 2589510.

Becker, G. (1968). "Crime and Punishment: an Economic Approach". *Journal of Political Economy*, 76:169–217.

Becker, G. (1974). "A Theory of Marriage". In Schultz, T. W., editor, *Economics of the Family*, pages 293–344. Chicago: University of Chicago Press.

Becker, G. (1993). "Nobel Lecture: The Economic Way of Looking at Behavior". *Journal of Political Economy*, 101(3):385–409.

Becker, G. (1996). *Accounting for Tastes*. Harvard University Press.

Benhabib, J. and Bisin, A. (2005). "Modelling Internal Commitment Mechanisms and Self-Control: a Neuroeconomics Approach to Consumption-Saving Decisions". *Games and Economic Behavior*, 52:460–492.

Bennett, M., Fingleton, J., Fletcher, A., Hurley, L., and Ruck, D. (2010). "What Does Behavioural Economics Mean for Competition Policy?". *Competition Policy International*, 6:111–137.

Berg, N. and Gigerenzer, G. (2010). "As-if Behavioral Economics: Neoclassical Economics in Disguise?". *History of Economic Ideas*, 18:133–166.

Berlin, I. (1958). *Four Essays on Liberty*. Oxford University Press.

Bernheim, B. D. (1986). Axiomatic Characterizations of Rational Choice in Strategic Environments. *Scandinavian Journal of Economics*, 88(3):473–488.

Bernheim, D. and Rangel, A. (2004). "Addiction and Cue-Triggered Decision Processes". *American Economic Review*, 94:1558–1590.

Bernheim, D. and Rangel, A. (2007). "Toward Choice-Theoretic Foundations for Behavioral Welfare Economics". *AEA Papers and Proceedings*, 97:464–470.

Bernheim, D. and Rangel, A. (2009). "Beyond Revealed Preferences: Choice-Theoretic Foundations for Behavioral Welfare Economics". *The Quarterly Journal of Economics*, 124(1):51–104.

Bester, H. and Güth, W. (1998). "Is Altruism Evolutionary Stable?". *Journal of Economic Behavior and Organisation*, 34:211–221.

Bicchieri, C. (2004). "Rationality and Game Theory". In *The Oxford Dictionary of Rationality*, pages 182–205. Oxford University Press.

Binmore, K. (1994). *Playing Fair: Game Theory and the Social Contract (vol. I)*. MIT Press.

Binmore, K. (1999). "Why Experiment in Economics". *The Economic Journal*, 109(453):F16–F24.

Binmore, K. (2010). "Social Norms or Social Preferences?". *Mind and Society*, 9:139–158.

Binmore, K. and Shaked, A. (2010). "Experimental Economics: Where Next?". *Journal of Economic Behavior and Organization*, 73(1):87–100.

Bleichrodt, H., Pinto, J. L., and Wakker, P. (2001). "Making Descriptive Use of Prospect Theory to Improve the Prescriptive Use of Expected Utility". *Management science*, 47(11):1498–1514.

Bloch, F., Sánchez-Pagés, S., and Soubeyran, R. (2006). When does universal peace prevail? Secession and group formation in conflict. *Economics of Governance*.

Blomquist, W. (1994). "Changing Rules, Changing Games: Evidence from Groundwater Systems in Southern California". In Ostrom, E., Gardner, R., and Walker, J., editors, *Rules, Games and Common-Pool Resources*, pages 193–300. Ann Arbor: University of Michigan Press.

Boianovsky, M. (2013). "Before Macroeconomics: Pareto and the Dynamics of the Economic Aggregate". *Revue européenne des sciences sociales*, 51(2):103–131.

Bolle, F. (2000). "Is Altruism Evolutionary Stable? And Envy and Malevolence? - Remarks on Bester and Güth". *Journal of Economic Behavior and Organisation*, 42:131–133.

Bolton, G. and Ockenfels, A. (2000). "ERC: A Theory of Inequity, Reciprocity and Competition". *American Economic Review*, 90:166–193.

Bordalo, P., Gennaioli, N., and Schleifer, A. (2013). "Salience and Consumer Choice". *Journal of Political Economy*, 121:803–843.

Bordas, L. (1847). "Mesure de l'utilité des travaux publics; réponse à M. Dupuit". *Annales des Ponts et Chaussées*, 13:247–284.

Boudon, R. (1974). *Education, Opportunity, and Social Inequality: Changing Prospects in Western Society*. Wiley-Interscience, New York.

Boudon, R. (1994). *The Art of Self-Persuasion: the Social Explanation of False Beliefs*. Polity Press, Cambridge.

Boumans, M. (2004). *How Economists Model the World into Numbers*. London: Routledge.

Bourdieu, P. (1974). "Avenir de classe et causalité du probable". *Revue franÃğaise de sociologie*, 15:3–42.

Bovens, L. (2009). "The Ethics of Nudge". In Grüne-Yanoff, T. and Hansson, S., editors, *Preference Change: Approaches from Philosophy, Economics and Psychology*, pages 207–219. Berlin and New York: Springer.

Boyd, R. and Richerson, P. (1985). *Culture and the Evolutionary Process*. University of Chicago Press.

Bratman, M. (1993). "Shared Intentions". *Ethics*, 104:97–113.

Breen, R. and Goldthorpe, J. H. (1997). "Explaining Educational Differentials: Towards a Formal Rational Action Theory". *Rationality and Society*, 9(3):275–303.

Brocas, I. and Carrillo, J. (2008). "The Brain as a Hierarchical Organization". *American Economic Review*, 98:1312–1346.

Broome, J. (1991a). "Utility". *Economics and Philosophy*, 7:1–12.

Broome, J. (1991b). *Weighing Goods: Equality, Uncertainty, and Time*. Oxford: Basil Blackwell.

Bruni, L. and Sugden, R. (2007). "The Road Not Taken: How Psychology Was Removed From Economics, And How It Might Be Brought Back". *The Economic Journal*, 117(516):146–173.

Bruni, L. and Sugden, R. (2008). "Fraternity: Why the Market Need Not Be a Morally Free Zone". *Economics and Philosophy*, 24(1):35–64.

Buchanan, J. (1968). *The Demand and Supply of Public Goods*. Chicago: Rand McNally.

Buss, S. (2014). "Personal Autonomy". In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2014 edition.

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

Camerer, C., Issacharoff, S., Loewenstein, G., O'Donoghue, T., and Rabin, M. (2003). "Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism' ". *University of Pennsylvania Law Review*, 151:1211–1254.

Cardenas, J., Stranlund, J., and Willis, C. (2000). "Local Environmental Control and Institutional Crowding-out". *World Development*, 28(10):1719–1733.

Carpenter, J. P. and Matthews, P. H. (2003). "Beliefs, Intentions, and Evolution: Old versus New Psychological Game Theory". *Behavioral and Brain Sciences*, 26:158–159.

Carrasco, M. (2011). "Hutcheson, Smith and Utilitarianism". *The Review of Metaphysics*, 64(3):515–553.

Casajus, A. (2000). "Focal Points in Framed Strategic Forms". *Games and Economic Behavior*, 32:263–291.

Chander, P. and Tulkens, H. (1997). "The Core of an Economy with Multilateral Externalities". *International Journal of Game Theory*, 26:379–401.

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

Choi, J., Laibson, D., Madrian, B., and Metrick, A. (2004). "For Better or for Worse: Default Effects and 401(k) Savings Behavior". In Wise, A., editor, *Perspectives on the Economics of Aging*. Chicago: Univ. Chicago Press.

Christman, J. (1991). "Autonomy and Personal History". *Canadian Journal of Philosophy*, 21:1–24.

Christman, J. (1993). "Defending Historical Autonomy: A Reply to Professor Mele". *Canadian Journal of Philosophy*, 23:281–290.

Cohen, G. (1989). "On the Currency of Egalitarian Justice". *Ethics*, 99(4):906–944.

Colman, A. (2003). "Cooperation, Psychological Game Theory, and Limitations of Rationality in Social Interaction". *Behavioral and brain sciences*, 26:139–198.

Colman, A. and Bacharach, M. (1997). "Payoff Dominance and the Stackelberg Heuristic". *Theory and Decision*, 43:1–19.

Conly, S. (2013). *Against Autonomy. Justifying Coercive Paternalism*. Cambridge: Cambridge University Press.

Cournot, A.-A. (1838). *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette.

Cox, M., Arnold, G., and Tomás, S. (2010). "A Review of Design Principles for Community-based Natural Resource Management". *Ecology and Society*, 15(4).

Crawford, V. (2013). "Boundedly Rational versus Optimization-Based Models of Strategic Thinking and Learning in Games". *Journal of Economic Literature*, 51(2):512–527.

Crawford, V., Costa-Gomes, M., and Iriberri, N. (2013). "Structural Models of Non-equilibrium Strategic Thinking: Theory, Evidence, and Applications". *Journal of Economic Literature*, 51(1):5–62.

Cubitt, R. and Sugden, R. (2001). "On Money Pumps". *Games and Economic Behavior*, 37:121–160.

Currarini, S. and Marini, M. (1998). "The Core of Games with Stackelberg Leaders". MPRA Paper 2219, University Library of Munich.

Cuypers, S. and Ishtiyague, H. (2008). "Educating for Well-Being and Autonomy". *Theory and Research in Education*, 6(1):71–93.

Darwall, S. (2003). *Contractarianism, Contractualism*. Wiley-Blackwell.

Darwall, S. (2006a). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Harvard University Press.

Darwall, S. (2006b). "The Value of Autonomy and the Autonomy of the Will". *Ethics*, 116:263–284.

d'Aspremont, C., Jacquemin, J., Gabszeweiz, J., and Weymark, J. (1983). "On the Stability of Collusive Price Leadership". *Canadian Journal of Economics*, 16:17–25.

Davis, J. (2011). *Individual and Identity in Economics*. Cambridge University Press.

Debreu, G. (1952). "A Social Equilibrium Existence Theorem". In *National Academy of Sciences of the USA*, volume 38.

Demeulenaere, P. (1996). *Homo oeconomicus. Enquete sur la constitution d'un paradigme*. PUF.

Dietrich, F. and List, C. (2013a). "A Reason-Based Theory of Rational Choice". *Noûs*, 47(1):104–134.

Dietrich, F. and List, C. (2013b). "Reasons for (Prior) Belief in Bayesian Epistemology". *Synthese*, 190(5):787–808.

Dietrich, F. and List, C. (2013c). "Where Do Preferences Come From?". *International Journal of Game Theory*, 42(3):613–637.

Dietrich, F. and List, C. (2014). "Reason-Based Rationalization". unpublished manuscript.

Dubois, D., Willinger, M., and Van Nguyen, P. (2012). Optimization incentive ane relative riskiness in experimental stag hunt games. *International Journal of Game Theory*, 41(2):369–380.

Dufwenberg, M. and Kirchsteiger, G. (2004). "A Theory of Sequential Reciprocity". *Games and Economic Behaviour*, 47:268–298.

Dupuit, J. (1844). "De la mesure de l'utilité des travaux publics". *Annales des Ponts et Chaussées*, 8:332–375.

Dupuit, J. (1849). "De l'influence des péages sur l'utilité des voies de communication". *Annales des Ponts et Chaussées*, 17(1):170–248.

Dworkin, G. (1972). "Paternalism". *Monist*, 56:64–84.

Eckel, C. and Gintis, H. (2010). "Blaming the Messenger: Notes on the Current State of Experimental Economics". *Journal of Economic Behavior and Organization*, 73(1):109–119.

Edgeworth, F. Y. (1881). *Mathematical Psychics: an Essay on the Application of Mathematics to the Moral Sciences*. New-York: Augustus M. Kelley.

Ekelund, R. and Hébert, R. (1999). *Secret Origins of Modern Microeconomics: Dupuit and the Engineers*. University of Chicago Press.

Ellet, C. J. (1839). *An Essay on the Laws of Trade in Reference to the Works of Internal Improvement in the United States*. New York: Augustus M. Kelley. Reprint 1966.

Elster, J. (1986). Introduction. In Elster, J., editor, *Rational Choice*, pages 1–33. New York University Press.

Elster, J. (1989). *The Cement of Society*. Cambridge: Cambridge University Press.

Falk, A., Fehr, E., and Fischbacher, U. (2008). "Testing Theories of Fairness — Intentions Matter". *Games and Economic Behavior*, 62:287–303.

Falk, A. and Fischbacher, U. (2006). "A Theory of Reciprocity". *Games and Economic Behavior*, 54(2):293–315.

Fehr, E. and Schmidt, K. (1999). "A Theory of Fairness, Competition and Cooperation". *Quaterly Journal of Economics*, 114:817–868.

Fehr, E. and Schmidt, K. (2010). "On Inequity Aversion: A Reply to Binmore and Shaked". *Journal of Economic Behavior and Organization*, 73(1):101–108.

Feinberg, J. (1971). "Legal Paternalism". *Canadian Journal of Philosophy*, 1:106–124.

Feinberg, J. (1986). *Harm to Self: The Moral Limits of the Criminal Law*. Oxford University Press.

Fershtman, C. and Gneezy, U. (2001). "Strategic Delegation: an Experiment". *RAND Journal of Economics*, 32(2):352–368.

Fershtman, C. and Kalai, E. (1997). "Unobserved Delegation". *International Economic Review*, 38(4):763–774.

Fischer, J. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Franck, R. (1987). "If Homo economicus Could Choose his Own Utility Function, Would he Choose One With a Conscience?". *American Economic Review*, 77(4):593–604.

Franck, R. (1988). *Passions within Reason – The Strategic Role of the Emotions*. New York: W.W. Norton.

Frederick, S. (2003). "Time Preference and Personal Identity". In Loewenstein, G., Read, D., and Baumeister, R., editors, *Time and Decision: Psychological Perspectives in Intertemporal Choice*. New York: Russel Sage.

Frederick, S., Loewenstein, G., and O'Donoghue, T. (2002). "Time Discounting and Time Preference: A Critical Review". *Journal of Economic Literature*, 40(2):351–401.

Frey, B. and Oberholzer-Gee, F. (1997). "The Cost of Price Incentives : An Empirical Analysis of Motivation Crowding-Out". *American Economic Review*, 87(4):746–755.

Frey, B. and Stutzer, A. (2002). "What Can Economists Learn from Happiness Research?". *Journal of Economic Literature*, 40:402–435.

Friedman, M. (1953). *Essays in Positive Economics Part I - The Methodology of Positive Economics*. University of Chicago Press.

Friedman, M. and Savage, L. (1948). "The Utility Analysis of Choices Involving Risk". *Journal of Political Economy*, 56.

Fudenberg, D. and Levine, D. (2006). "A Dual-Self Model of Impulse Control". *American Economic Review*, 96:1449–1476.

Gauthier, D. (1969). *The Logic of Leviathan: The Moral and Political Theory of Thomas Hobbes*. Oxford: Clarendon Press.

Gauthier, D. (1986). *Morals by Agreement*. Oxford: Oxford University Press.

Gauthier, D. (1991). "Why Contractarianism?". In Vallentyne, P., editor, *Contractarianism and Rational Choice*, pages 15–30. Cambridge: Cambridge University Press.

Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989). "Psychological Games and Sequential Rationality". *Games and Economic Behavior*, 1:60–79.

Georgescu-Roegen, N. (1971). *The Entropy Law and the Economic Process*. Cambridge (Mass.): Harvard University Press.

Gigerenzer, G. and Selten, R., editors (2001). *Bounded Rationality: the Adaptive Toolbox*. Cambridge: MIT Press.

Gigerenzer, G., Todd, P., and the ABC Research Group (1999). *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Gilbert, M. (1989). *On Social Facts*. London: Routledge.

Gilbert, M. (1997). "What is it for us to Intend?". In Holmstrom-Hintikka, G. and Tuomela, R., editors, *Contemporary Action Theory*, volume 2, pages 65–85. Dordrecht: Springer.

Gintis, H. (2009). *The Bounds of Reason*. Princeton University Press.

Gloria-Palermo, S. (1999). *The Evolution of Austrian Economics*. Routledge Studies in the History of Economics.

Gold, N. and Sugden, R. (2007). "Theories of Team Agency". In Peter, F. and Schmid, H. B., editors, *Rationality and Commitment*, pages 280–312. Oxford: Oxford University Press.

Goldman, A. (1989). "Interpretation Psychologized". *Mind and Language*, 4:161–189.

Goldman, A. (1992). "In Defense of the Simulation Theory". *Mind and Language*, 7:104–119.

Goldstein, D. and Gigerenzer, G. (2002). "Models of Ecological Rationality: the Recognition Heuristic". *Psychological Review*, 109:75–90.

Gordon, R. (1986). "Folk Psychology as Simulation". *Mind and Language*, 1:158–171.

Gordon, R. (1992). "The Simulation Theory: Objections and Misconceptions". *Mind and Language*, 7:11–34.

Grant, R. (2002). "The Ethics of Incentives: Historical Origins and Contemporary Understandings". *Economics and Philosophy*, 18(01):111–139.

Grüne-Yanoff, T. (2011). "Evolutionary Game Theory, Interpersonal Comparisons and Natural Selection: A Dilemma". *Philosophy and Biology*, 26:637–654.

Grüne-Yanoff, T. (2012). "Old Wine in New Casks: Libertarian Paternalism Still Violates Liberal Principles". *Social Choice and Welfare*, 38(4):635–645.

Guala, F., Mittone, J., and Ploner, M. (2013). "Group Membership, Team Preferences, and Expectations". *Journal of Economic Behavior and Organization*, 86:183–190.

Güth, W. (1995). "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives". *International Journal of Game Theory*, pages 323–344.

Güth, W. and Yaari, M. (1992). "Explaining Reciprocal Behavior in Simple Strategic Games: an Evolutionary Approach". In Witt, U., editor, *Explaining Forces and Changes: Approaches to Evolutionary Economics*. University of Michigan Press.

Gutmann, A. (1987). *Democratic Education*. Princeton: Princeton University Press.

Halpern, D. and Nesterak, M. (2014). "Nudging in the UK: a Conversation with David Halpern". The Psych Report. http://thepsychreport.com/conversations/nudging-the-uk-a-conversation-with-david-halpern/.

Hammerstein, P. and Selten, R. (1994). "Game Theory and Evolutionary Biology". In Aumann, R. and Hart, S., editors, *Handbook of Game Theory*, pages 929–993. North Holland.

Hands, D. W. (2010). "The Positive-Normative Dichotomy and Economics". In Mäki, U., editor, *Philosophy of Economics*. Amsterdam: Elsevier. Vol. 13 of D. Gabbay, P. Thagard, and J. Woods (Eds.), *Handbook of the philosophy of science*.

Hansen, G. and Jespersen, M. (2013). "Nudge and the Manipulation of Choice: a Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy". *European Journal of Risk Regulation*, 1:3–28.

Harsanyi, J. (1955). "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility". *Journal of Political Economy*, 63:309–321.

Harsanyi, J. and Selten, R. (1988). *A General Theory of Equilibrium Selection in Games*. Cambridge: MIT Press.

Harstad, R. and Selten, R. (2013). "Bounded-Rationality Models: Tasks to Become Intellectually Competitive". *Journal of Economic Literature*, 51(2):496–511.

Hausman, D. (2012). *Preference, Value, Choice, and Welfare*. Cambridge University Press.

Hausman, D. (2015). "On the Econ Within". unpublished manuscript.

Hausman, D. and Welch, B. (2010). "Debate: To Nudge or Not To Nudge". *Journal of Political Philosophy*, 18:123–136.

Hébert, R. F. (1998). "Jevons and Menger Re-homogenized : Who Is the Real 'Odd Man Out'? A Comment on Peart". *American Journal of Economics and Sociology*, 57(3):327–332.

Heifetz, A., Shannon, C., and Spiegel, Y. (2007a). "The Dynamic Evolution of Preferences". *Economic Theory*, 32:251–286.

Heifetz, A., Shannon, C., and Spiegel, Y. (2007b). "What to Maximize if You Must". *Journal of Economic Theory*, 133:31–57.

Hobbes, T. (1651). *Leviathan, or The Matter, Forme and Power of a Common Wealth Ecclesiasticall and Civill*. McMaster University Archive of the History of Economic Thought.

Hodgson, D. (1967). *Consequences of Utilitarianism*. Oxford: Oxford Clarendon Press.

Hollis, M. (1998). *Trust within Reason*. Cambridge: Cambridge University Press.

Huck, S. and Oechssler, J. (1998). "The Indirect Evolutionary Approach to Explaining Fair Allocations". *Games and Economic Behavior*, 28:13–24.

Hume, D. (1739). *A Treatise of Human Nature*. Oxford: Oxford University Press. reprint 2000.

Hurley, S. (1989). *Natural Reasons*. Oxford: Oxford University Press.

Infante, G., Lecouteux, G., and Sugden, R. (2015). "Preference Purification and the Inner Rational Agent: a Critique of the Conventional Wisdom of Behavioural Welfare Economics". *Journal of Economic Methodology*. forthcoming.

Jaffé, W. (1965). *Correspondence of Léon Walras and related papers*, volume I. North Holland Publishing.

Jaffé, W. (1976). "Menger, Jevons and Walras De-Homogenized". *Economic Inquiry*, 14(4):511–524.

Janssen, M. (2001). "Rationalizing Focal Points". *Theory and Decision*, 50:119–148.

Janssen, M. (2006). "On the Strategic Use of Focal Points in Bargaining Situations". *Journal of Economic Psychology*, 27:622–634.

Jevons, W. (1866). "Brief Account of a General Mathematical Theory of Political Economy". *The Journal of the Royal Statistical Society*, pages 282–287.

Jevons, W. (1871). *Theory of Political Economy*. New York: Augustus M. Kelley, 5th edition (1965) edition.

Jolink, A. (2005). *Evolutionist Economics of Leon Walras*. Routledge.

Kahneman, D. (1996). "Comment on Plott (1996)". In Arrow, K., Colombatto, E., Perlman, M., and Schmidt, C., editors, *The Rational Foundations of Economic Behaviour*, pages 251–254. Basingstoke: International Economic Association and Macmillan.

Kahneman, D. (2003). "A Perspective on Judgment and Choice: Mapping Bounded Rationality". *American Psychologist*, 58:697–720.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., Knetsch, J., and Thaler, R. (1990). "Experimental Tests of the Endowment Effect and the Coase Theorem". *Journal of Political Economy*, 98:1325–1348.

Kahneman, D. and Tversky, A. (1979). "Prospect Theory: An Analysis of Decision under Risk". *Econometrica*, 47(2):263–291.

Kahneman, D. and Tversky, A., editors (2000). *Choice, Value, and Frames*. Cambridge: Cambrdige University Press.

Kahneman, D., Wakker, P., and Sarin, R. (1997). "Back to Bentham? Explorations of Experienced Utility. *Quaterly Journal of Economics*, 112:375–405.

Kamii, C. (1991). "Toward Autonomy: The Importance of Critical Thinking and Choice Making". *School Psychology Review*, 20(3):382–388.

Kandori, M., Mailath, G., and Rob, R. (1993). "Learning, Mutation, and Long Run Equilibria in Games". *Econometrica*, 61(1):29–56.

Kant, I. (1785). *Groundwork of the Metaphysics of Morals*. Cambridge: Cambridge Univsity Press. Translated by Mary Gregor (1997).

Kant, I. (1797). *Metaphysics of Morals*. Cambridge: Cambridge Univsity Press. Translated by Mary Gregor (1996).

Kavka, G. (1983). "The Toxin Puzzle". *Analysis*, 43:33–36.

Kockesen, L., Ok, E., and Sethi, R. (2000a). "Evolution of Interdependent Preferences in Aggregative Games". *Games and Economic Behavior*, 31:303–310.

Kockesen, L., Ok, E., and Sethi, R. (2000b). "The Strategic Advantage of Negatively Interdependent Preferences". *Journal of Economic Theory*, 92:274–299.

Korsgaard, C. (1989). "Personal Identity and the Unity of Agency: A Kantian Response to Parfit". *Philosophy and Public Affairs*, 18(2):101–132.

Korsgaard, C. (2009). *Self-Constitution, Agency, Identity, and Integrity*. Oxford: Oxford University Press.

Köszegi, B. and Rabin, M. (2007). "Mistakes in Choice-Based Welfare Analysis". *AEA Papers and Proceedings*.

Köszegi, B. and Rabin, M. (2008). "Choices, Situations, and Happiness". *Journal of Public Economics*, 92(8-9):1821–1832.

Lardner, D. (1850). *Railway Economy*. New York: Augustus M. Kelley. Reprint 1968.

Larrouy, L. (2013). "Bacharach's 'Variable Frame Theory': A Legacy from Schelling's Issue in the Refinement Program?". GREDEG Working Paper 2013-11.

Layard, R. (2005). *Happiness: Lessons From a New Science*. London: Allen Lane.

Lecouteux, G. (2015). "In Search of Lost Nudges". *The Review of Philosophy and Psychology*. DOI:10.1007/s13164-015-0265-0.

Lecouteux, G. and Moulin, L. (2015). "To Gain or Not to Lose? Tuition Fees for Loss Averse Students". *Economics Bulletin*, 35(2):1005–1019.

Levine, D. (2012). *Is Behavioral Economics Doomed? The Ordinary versus the Extraordinary*. OpenBook Publishers.

Levitt, S. and List, J. (2008). "Homo economicus Evolves". *Science*, 319(5865):909–910.

Li, C., Li, Z., and Wakker, P. (2014). "If Nudge Cannot be Applied: a Litmus Test of the Readers' Stance on Paternalism". *Theory and Decision*, 76:297–315.

Liberman, N., Trope, Y., and Stephan, E. (2007). Psychological distance. *Social psychology: Handbook of basic principles*, 2:353–383.

Lindhal, E. (1919). "Just Taxation – A Positive Solution". In Musgrave, R. and Peacock, A., editors, *Classics in the Theory of Public Finance (1958)*. London: McMillan.

Lipsey, R. and Lancaster, K. (1956). "The General Theory of Second Best". *Review of Economic Studies*, 24:11–32.

Loewenstein, G. and O'Donoghue, T. (2004). "Animal Spirits: Active and Deliberative Processes in Economic Behavior". unpublished manuscript.

Loewenstein, G., Rick, S., and Cohen, J. D. (2008). "Neuroeconomics". *Annual Review of Psychology*, 59:647–72.

Loomes, G., Starmer, C., and Sugden, R. (2003). "Do Anomalies Disappear in Repeated Markets?". *The Economic Journal*, 113(486):C153–C166.

Lucas, R. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester conference series on public policy*.

Mäki, U. (2003). "'The methodology of positive economics' (1953) does not give us *the* methodology of positive economics". *Journal of Economic Methodology*, 10(4):495–505.

Mariotti, M. (1995). "Is Bayesian Rationality Incompatible with Strategic Rationality?". *Economic Journal*, 105:1099–1109.

Marshall, A. (1890). *Principles of Economics*. London McMillan, 8th edition edition. reprint 1966.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge University Press.

Maynard Smith, J. and Price, G. (1973). "The Logic of Animal Conflict". *Nature*, 246:15–18.

McQuillin, B. and Sugden, R. (2012a). "How the Market Responds to Dynamically Inconsistent Preferences". *Social Choice and Welfare*, 38(4):617–634.

McQuillin, B. and Sugden, R. (2012b). "Reconciling Normative and Behavioural Economics: the Problems to be Solved". *Social Choice and Welfare*, 38:553–567.

Mehta, J., Starmer, C., and Sugden, R. (1994a). "Focal Points in Pure Coordination Games: An Experimental Investigation". *Theory and Decision*, 36:163–185.

Mehta, J., Starmer, C., and Sugden, R. (1994b). "The Nature of Salience: An Experimental Investigation of Pure Coordination Games". *American Economic Review*, 84:658–673.

Mele, A. (1993). "History and Personal Autonomy". *Canadian Journal of Philosophy*, 23:271–280.

Mele, A. (1995). *Autonomous Agents: From Self-Control to Autonomy*. New York: Oxford University Press.

Menger, C. (1871). *Grunsätze des Volkswirthschaftlehre*. Ludwig Von Mises Institute. Reprint 2007.

Menger, C. (1883). *Investigations Into the Method of the Social Sciences with Special References to Economics*. New York University Press. Reprint 1985.

Menger, K. (1973). "Austrian Marginalism and Mathematical Economics". In Hicks, J. and Weber, W., editors, *C. Menger and the Austrian School of Economics*, pages 54–60. Oxford: Clarendon Press.

Milgram, S. (1975). *Obedience to Authority*. New York: Harper Colophon.

Mill, J. (1859). *On Liberty*. Meridian Book. in *Utilitarianism, On Liberty, Essay on Bentham, together with seleceted writings of Jeremy Bentham and John Austin* (1974).

Mill, J. (1882). *A System of Logic, Ratiocinative and Inductive*. Harper & Brothers, 8th edition.

Mongin, P. and Cozic, M. (2014). "Rethinking Nudges". available ar SSRN: http://ssrn.com/abstract=2529910.

Nagatsu, M. (2015). "Social Nudges: Their Mechanisms and Justification". *The Review of Philosophy and Psychology*. forthcoming.

Nagel, T. (1986). *The View from Nowhere*. Oxford: Oxford University Press.

Nash, J. (1950). "The Bargaining Problem". *Econometrica*, 18:155–162.

Nelkin, D. (2007). "Do We Have a Coherent Set of Intuitions about Moral Responsibility?". *Midwest Studies in Philosophy*, 31:243–259.

Nussbaum, M. (2000). *Women and Human Development*. Cambridge: Cambridge University Press.

O'Donoghue, T. and Rabin, M. (1999). "Doing It Now or Later". *American Economic Review*, 89(1):103–124.

Ostman, A. (1998). "External Control May Destroy the Commons". *Rationality and Society*, 10(1):103–122.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. New-York: Cambridge University Press.

Ostrom, E., Gardner, R., and Walker, J. (1994). *Rules, Games and Common-Pool Resources*. Ann Arbor: University of Michigan Press.

Pantaleoni, M. (1889). *Pure Economics*. London: McMillan. Translated by T. B. Bruce (1898).

Pareto, V. (1909). *Manual of Political Economy*. London: McMillan. Translated by A. Schwier from the 1927 french edition (1971).

Pareto, V. (1916). *The Mind and Society: a Treatise on General Sociology*. London: Jonathan Cape. Translated by A. Bongiorno and A. Livingston (1936).

Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.

Parfit, D. (2010). *On What Matters*. Oxford University Press.

Peart, S. (1998). "Jevons and Menger Re-homogenized? Jaffé After 20 Years". *American Journal of Economics and Sociology*, 57(3):307–325.

Pigou, A. (1920). *The Economics of Welfare*. London: McMillan.

Plott, C. (1996). "Rational Individual Behaviour in Markets and Social Choice Processes: the Discovered Preference Hypothesis". In Arrow, K., Colombatto, E., Perlman, M., and Schmidt, C., editors, *The Rational Foundations of Economic Behaviour*, pages 225–250. Basingstoke: International Economic Association and Macmillan.

Possajennikov, A. (2000). "On the Evolutionary Stability of Altruistic and Spiteful Preferences". *Journal of Economic Behavior and Organization*, 42:125–129.

Poulsen, A. U. and Roos, M. W. (2012). "Do People Make Strategic Commitments? Experimental Evidence on Strategic Information Avoidance". *Experimental Economics*, 13:206–225.

Qizilbash, M. (2011). "Sugden's Critique of Sen's Capability Approach and the dangers of libertarian paternalism". *International Review of Economics*, 58(1):21–42.

Qizilbash, M. (2012). "Informed Desire and the Ambitions of Libertarian Paternalism". *Social Choice and Welfare*, 38:647–658.

Rabin, M. (1993). "Incorporating Fairness into Game Theory and Economics". *American Economic Review*, 83:1281–1302.

Rabin, M. (2013). "Incorporating Limited Rationality into Economics". *Journal of Economic Literature*, 51(2):528–543.

Rawls, J. (1971). *A Theory of Justice*. Cambridge: Harvard University Press.

Read, D. and van Leeuwen, B. (1998). "Predicting Hunger: the Effects of Appetite and Delay on Choice". *Organizational Behavior and Human Decision Processes*, 76:189–205.

Regan, D. (1980). *Utilitarianism and Cooperation*. Oxford: Oxford Clarendon Press.

Roemer, J. (1998). *Equality of Opportunity*. Cambridge, Mass: Harvard University Press.

Roth, A. (1991). "Game Theory as a Aart of Empirical Economics". *The Economic Journal*.

Roth, A. (2007). "Repugnance as a Constraint on Markets". *Journal of Economic Perspectives*, 21(3):37–58.

Roth, A. (2012). "The Theory and Practice of Market Design". In *The Nobel Prize*, pages 343–363.

Rousseau, J.-J. (1762). *Du Contrat Social*. Paris: LGF. Reprint 1996.

Rubinstein, A. and Salant, Y. (2012). "Eliciting Welfare Preferences from Behavioural Data Sets". *Reveiw of Economics Studies*, 79:375–387.

Ruhm, C. (2012). "Understanding Overeating anf Obesity". *Journal of Health Economics*, 31(6):781–796.

Salant, Y. and Rubinstein, A. (2008). "(A,f): Choice with Frames". *Review of Economic Studies*, 75:1287–1296.

Sally, D. (1995). "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992". *Rationality and Society*, 7(1):58–92.

Samuelson, L. (2001). "Introduction to the Evolution of Preferences". *Journal of Economic Theory*, 97(2):225–230.

Samuelson, P. (1937). "A Note on the Measurement of Utility". *Review of Economic Studies*, 4:155–161.

Samuelson, P. (1947). *Foundations of Economic Analysis*. Harvard University Press.

Savage, L. (1954). *The Foundations of Statistics*. New York: Wiley.

Schelling, T. (1960). *The Strategy of Conflict*. Cambridge, Mass.: Harvard University Press.

Schelling, T. (1978). *Micromotives and Macrobehavior*. W. W. Norton and Company. Reprint 2006.

Schkade, D. and Kahneman, D. (1998). "Does Living in California Make People Happy? A Focusing Illusion in Judgments of Life Satisfaction". *Psychological Science*, 9:340–346.

Schmidt, D., Shupp, R., Walker, J., and Ostrom, E. (2003). "Playing Sage in Coordination Games: the Roles of Risk Dominance, Payoff Dominance, and History of Play". *Games and Economic Behavior*, 42:281–299.

Searle, J. (1990). "Collective Intentions and Actions". In Cohen, P., Morgan, J., and Pollack, M., editors, *Intentions in Communication*, pages 401–15. Cambridge, Mass.: MIT Press.

Sen, A. (1977). "Rational Fools : A Critique of the Behavioral Foundations of Economic Theory". *Philosophy and Public Affairs*, 6(4):317–344.

Sen, A. (1998). *Inequality Reexamined*. Cambridge, Mass: Harvard University Press.

Sen, A. (1999). *Development as Freedom*. Oxford: Oxford University Press.

Sen, A. (2002). *Rationality and Freedom*. Cambridge, Mass.: Belknap Press.

Sen, A. (2004). "Capabilities, Lists and Public Reason: Continuing the Conversation". *Feminist Economics*, 10(3):77–80.

Sen, A. (2009). *The Idea of Justice*. Cambridge MA: The Belknap Press of Harvard University Press.

Sengul, M., Gimeno, J., and Dial, J. (2012). "Strategic Delegation: A Review, Theoretical Integration, and Research Agenda". *Journal of Management*, 38(1):375–414.

Shiffrin, S. (2000). "Paternalism, Unconscionability Doctrine, and Accommodation". *Philosophy and Public Affairs*, 29:205–250.

Sidgwick, H. (1874). *The Methods of Ethics*. Hackett Publishing. Reprint 1981.

Simon, H. (1955). "A Behavioral Model of Rational Choice". *Quaterly Journal of Economics*, 66:99–118.

Simon, H. (1956). "Rational Choice and the Structure of the Environment". *Psychological Review*, 63:129–138.

Smerilli, A. (2012). "We-thinking and Vacillation between Frames: Filling a Gap in Bacharach's Theory". *Theory and Decision*, 73(4):539–60.

Smith, A. (1759). *The Theory of Moral Sentiments*. Penguin. Reprint 2010.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: Pickering and Chatto (Publishers) Limited. Reprint 1805.

Smith, V. (1989). "Theory, Experiment and Economics". *Journal of Economic Perspectives*, 3:151–169.

Smith, V. (1994). "Economics in the Laboratory". *Journal of Economic Perspectives*, 8(1):113–131.

Smith, V. (2003). "Constructivist and Ecological Rationality in Economics". *American Economic Review*, 93(3):465–508.

Smith, V. (2008). *Rationality in Economics: Constructivist and Ecological Forms*. Cambridge: Cambridge University Press.

Sobel, J. (2005). "Interdependent Preferences and Reciprocity". *Journal of Economic Literature*, 43(2):392–436.

Sober, E. and Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, Mass.: Harvard University Press.

Sugden, R. (1991). "Rational Choice: A Survey of Contributions from Economics and Philosophy". *The Economic Journal*, 101(407):751–785.

Sugden, R. (1993). "Thinking as a Team: Toward an Explanation of Nonselfish Behavior". *Social Philosophy and Policy*, 10:69–89.

Sugden, R. (1995). "A Theory of Focal Points". *Economic Journal*, 105:533–50.

Sugden, R. (2000). "Team Preferences". *Economics and Philosophy*, 16:175–204.

Sugden, R. (2001). "The Evolutionary Turn in Game Theory". *Journal of Economic Methodology*, 8(1):113–130.

Sugden, R. (2003). "The Logic of Team Reasoning". *Philosophical Explorations*, 6:165–181.

Sugden, R. (2004). "The Opportunity Criterion: Consumer Sovereignty Without the Assumption of Coherent Preferences". *American Economic Review*, 94(4):1014–1033.

Sugden, R. (2006). "Hume's Non-Instrumental and Non-Propositional Decision Theory". *Economics and Philosophy*, 22:365–391.

Sugden, R. (2007). "The Value of Opportunities Over Time when Preferences are Unstable". *Social Choice and Welfare*, 29(4):665–682.

Sugden, R. (2008). "Why Incoherent Preferences do not Justify Paternalism". *Constitutional Political Economy*, 19(3):226–248.

Sugden, R. (2011). "Mutual Advantage, Conventions and Team Reasoning". *International Review of Economics*, 58(1):9–20.

Sugden, R. (2013). "The Behavioural Economist and the Social Planner: to Whom Should Behavioural Welfare Economics be Addressed?". *Inquiry*, 56:519–538.

Sunstein, C. (2014a). "The Ethics of Nudging". Available at SSRN: http://ssrn.com/abstract=2526341.

Sunstein, C. (2014b). *Why Nudge? The Politics of Libertarian Paternalism.* Yale University Press.

Sunstein, C. (2015). "Nudges, Agency, Navigability, and Abstraction: A Reply to Critics". *The Review of Philosophy and Psychology.* forthcoming, preliminary version available at SSRN 2577018.

Sunstein, C. and Thaler, R. (2003). "Libertarian Paternalism Is Not an Oxymoron". *The University of Chicago Law Review*, 70(4):1159–1202.

Taylor, F. (1911). *The Principles of Scientific Management and Shop Management.* Routledge/Thoemmes Press. Reprint 1993.

Taylor, P. and Jonker, L. (1978). "Evolutionary Stable Strategies and Game Dynamics". *Mathematical Biosciences*, 40:145–156.

Thaler, R. and Benartzi, S. (2004). "Save More Tomorrow: Using Behavioral Economics to Increase Employee Savings". *Journal of Political Economy*, 110:S164–S187.

Thaler, R. and Sunstein, C. (2003). "Libertarian Paternalism". *AEA Papers and Proceedings*, 93(2):175–179.

Thaler, R. and Sunstein, C. (2008). *Nudge. Improving Decisions About Health, Wealth, and Happiness.* Yale University Press.

Titmuss, R. (1970). *The Gift Relationship: From Human Blood to Social Policy.* London: Allen and Unwin.

Tuomela, R. and Miller, K. (1988). "We-Intentions". *Philosophical Studies*, 53:367–89.

Tversky, A. and Kahneman, D. (1981). "The Framing of Decisions and the Psychology of Choice". *Science New Series*, 211(4481):453–458.

Tversky, A. and Kahneman, D. (1992). "Advances in Prospect Theory: Cumulative Representation of Uncertainty". *Journal of Risk and Uncertainty*, 5:297–323.

van Veelen, M. (2012). "Robustness Against Indirect Invasions". *Games and Economic Behavior*, 74:382–393.

Varoufakis, Y. (2015). "No Time for Games in Europe". *The New York Times.* http://www.nytimes.com/2015/02/17/opinion/yanis-varoufakis-no-time-for-games-in-europe.html?_r=0, accessed 08/04/2015.

Vickrey, W. (1961). "Counterspeculation, Auctions, and Competitive Sealed Tenders". *Journal of Finance*, 16(1):8–37.

von Mises, L. (1949). *Human Action. A Treatise on Economics*. Ludwig von Mises Institute. Reprint 1998.

von Stackelberg, H. (1934). *Marktform und Gleichgewicht*. Vienna, Berlin: Springer.

Walras, L. (1874). *Elements of Pure Economics*. George Allen & Unwin. Translated by W. Jaffé (1954).

Walras, L. (1896). *Études d'économie sociale, ou Théorie de la répartition de la richesse sociale*. Paris: Pichon et Durand-Auzias, Lausanne: Rouge.

Walras, L. (1898). *Études d'économie politique appliquée, ou Théorie de la production de la richesse sociale*. Paris: Pichon et Durand-Auzias, Lausanne: Rouge.

Wason, P. and Evans, J. (1975). "Dual processes in reasoning?". *Cognition*, 3:141–154.

Weibull, J. (1995). *Evolutionary Game Theory*. Cambrdige: MIT Press.

White, J. (1982). *The Aims of Education Restated*. Cambridge: Cambridge University Press.

Whitman, D. G. and Rizzo, M. J. (2015). "The Problematic Welfare Standards of Behavioral Paternalism". *The Review of Philosophy and Psychology*. forthcoming.

Wilson, C. and Waddams Price, C. (2010). "Do Consumers Switch to the Best Supplier?". *Oxford Economic Papers*, 62:647–668.

Wilson, T. and Gilbert, D. (2003). "Affective Forecasting". In Zanna, M., editor, *Advances in Experimental Social Psychology*, volume 35, pages 345–411. San Diego, CA: Elsevier.

Wolf, S. (1990). *Freedom within Reason*. New York: Oxford University Press.

Zizzo, D. (2004). "Positive Harmony Transformations and Equilibrium Selection in Two-Player Games". Oxford: Department of Economics, University of Oxford.

Zizzo, D. and Tan, J. (2003). "Game Harmony as a Predictor of Cooperation in 2X2 Games: An Experimental Study". Oxford: Department of Economics, University of Oxford.

**Abstract:** The aim of this thesis is to address from a methodological, philosophical and theoretical perspective the problem of how to reconcile normative and behavioural economics — the "reconciliation problem". The first part develops a methodological assessment of behavioural welfare economics (and more specifically of libertarian paternalism), which currently constitutes the most accepted approach to deal with the reconciliation problem. We argue that behavioural welfare economics is in the direct continuation of neoclassical welfare economics, and that its conception of individual behaviour is tightly connected to Pareto's reductionist model of the *Homo economicus*. It is assumed that an individual can be defined by "true" preferences — distinct from her revealed preferences — and that those preferences are consistent and context-independent. The satisfaction of those true preferences is taken as the normative criterion. The existence of true preferences requires accepting that people have access to a latent mode of reasoning able to generate consistent and context-independent preferences, and to treat individual decision-making as if it was the result of an optimisation problem faced by an inner rational agent trapped in an outer psychological shell. We argue on the contrary that we cannot define unambiguously such true preferences. We then argue that the normative challenge raised by behavioural economics is that individuals may lack of autonomy: rather than focusing on the satisfaction of one's preferences, we argue that what matters is the ability to choose one's own preferences. Since we suggest that the difficulties of behavioural welfare economics are related to its commitment to an implausible model of individual preferences, we provide in the second part of the thesis a model of preferences compatible with our methodological analysis: we take Bacharach's variable frame theory as the primitive of our analysis, and build a model of endogenous preferences, in which individuals can choose to some extent their own preferences. The main contributions of this model are that (i) it does not require the existence of latent true preferences, and (ii) the individuals are allowed to team reason. We assume that collective preferences are strategically chosen, and show that team reasoners tend to be aggressive (cooperative) with other teams in submodular (supermodular) games. We finally show that team reasoning can be empirically justified as an ecologically rational heuristic.

**Keywords:** behavioural welfare economics; libertarian paternalism; nudge; preferences; reconciliation problem; team reasoning.

**Résumé:** L'objet de cette thèse est d'offrir une analyse tant méthodologique que philosophique et théorique du "problème de réconciliation", i.e. des implications normatives des résultats de l'économie comportementale. La première partie vise à dresser une analyse méthodologique de l'économie du bien-être comportementale (et en particulier du paternalisme libertarien), qui constitue actuellement la principale approche pour traiter le problème de réconciliation. Nous montrons que l'économie du bien-être comportementale est dans la directe lignée de l'économie du bien-être néoclassique, et que le modèle de comportement individuel sur lequel elle repose est étroitement lié au réductionnisme du modèle de l'*Homo economicus* développé par Pareto. Il est supposé qu'un individu peut être défini par ses "vraies" préférences, distinctes de ses préférences révélées, et que ces préférences sont cohérentes et stables. Le critère normatif retenu est alors la satisfaction de ces vraies préférences. L'existence de telles vraies préférences requiert l'existence d'un mode de raisonnement latent chez les individus susceptible de générer des préférences cohérentes et stables, et ainsi représente le raisonnement individuel comme un problème d'optimisation pour un agent rationnel intérieur, séparé du monde par une enveloppe psychologique externe. Nous montrons qu'il n'est pas possible de définir de façon non ambiguë ce que seraient de telles vraies préférences. Nous suggérons ensuite que le problème normatif soulevé par les résultats de l'économie comportementale est que les individus sont susceptibles de manquer d'autonomie: au lieu de se concentrer sur la satisfaction de ses préférences, nous proposons comme critère normatif la capacité de choisir ses propres préférences. Dans la mesure où les difficultés de l'économie du bien-être comportementale sont liées à son modèle de préférences sous-jacent, nous proposons dans la seconde partie de cette thèse un modèle de préférences compatible avec notre analyse: nous prenons comme primitive de notre analyse la *variable frame theory* proposée par Bacharach, et développons un modèle de préférences endogènes, dans lequel les individus peuvent choisir dans une certaine mesure leurs propres préférences. Les principales contributions de ce modèle sont (i) l'absence de référence à une notion de vraies préférences sous-jacentes, et (ii) la possibilité de modéliser des phénomènes d'intentionnalité collective (*team reasoning*). Nous supposons que les préférences collectives des individus sont le résultat d'un choix stratégique, et montrons que les individus tendent à être volontairement agressif (coopératif) avec les autres groupes d'individus dans le cadre de jeux sous-modulaires (super-modulaires). Nous montrons également que le *team reasoning* peut être justifié empiriquement comme une heuristique écologiquement rationnelle.

**Mots clés:** économie du bien-être comportementale; paternalisme libertarien; nudge; préférences; intentionnalité collective