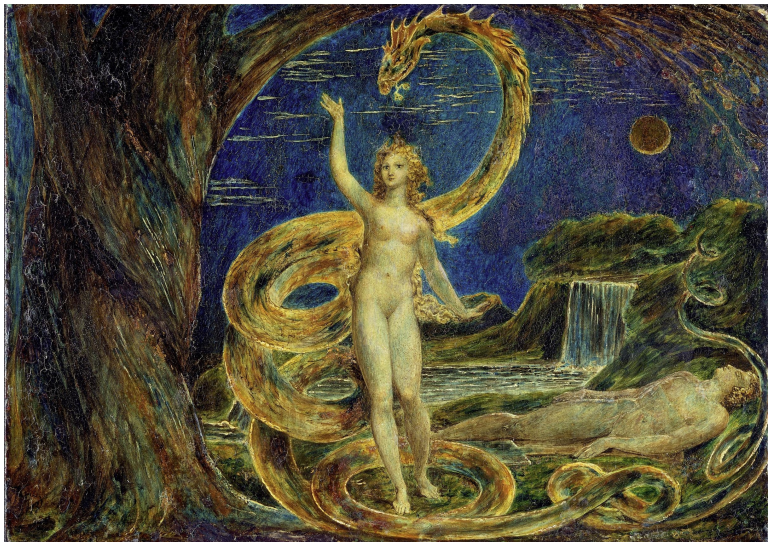
The London School of Economics
and Political Science

Promoting Social Norms via Microeconomics Teaching



Kamilla Haworth Buchter

A thesis submitted to the Department of Philosophy, Logic and Scientific Method of the London School of Economics and Political Science for the degree of Doctor of Philosophy, London, August, 2020

Front page photo credit

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of 84,632 words.

Statement of co-authored work

I confirm that Chapter 7 was jointly co-authored with Dr. Bjarke Mønsted and I contributed 50% of this work.

Abstract

In this thesis I argue that one way scientific descriptions can become self-fulfilling is by promoting social norms among the people they are disseminated to. Identifying this mechanism will enable us to change unwanted social implications caused by it. To make the argument, I rely on the definition of social norms given by Bicchieri [2006] in *The Grammar of Society* and use the case of microeconomics as it is presented in university textbooks. Thus, the aim of the thesis is to argue that one way microeconomics can be self-fulfilling is by promoting a social norm of self-interest - and often narrow self-interest - via its textbooks and university teaching practices.

To do this, I first use the current empirical findings to argue that the dissemination of the rationality assumption as it is presented in microeconomics textbooks can make microeconomics self-fulfilling. Second, I conduct a historical analysis to show that the claims that greed and self-interest are beneficial have been a part of modern economics from its beginning and still is today. I then discuss why the rationality assumption is a part of contemporary microeconomics and analyse how it is presented in standard textbook models today. Here, we see that even though some of the models can account for other-regarding preferences, the textbooks do not mention this fact. Instead, they present the rationality assumption as focusing on self-interested preferences only, and justify it as being both descriptively plausible and normatively desirable. Finally, I use the above analyses to argue that microeconomics textbooks and teaching practices can change people's behaviour by making them follow a social norm of self-interest in economic situations. I end the thesis by presenting the results of an empirical study designed to test this argument.

Acknowledgements

First and foremost, I would like to thank my supervisor Jason Alexander for his continuous help and perceptive comments throughout the programme. This thesis would not have been possible without your support.

Also, I want to thank the people at LSE with whom I have discussed my work. These include (but are not limited to) my secondary supervisors Christian List and Johanna Thoma, Roman Frigg, Richard Bradley, Gabriel Wollner, Mary Morgan, Dominic Berry, James Nguyen, Camilla Colombo, Silvia Milano, Nicolas Wütrich, and Catherine Greene.

I am grateful to Klemens Kappel for making my external stay at Copenhagen University possible. I also want to thank H. Orri Stefánsson, Thor Grünbaum, Nana Cecilie Halmsted Kongsholm, Georgiana Turculet, and Katla Heðinsdóttir for their helpful comments and support during this stay.

A special thanks should also go to my examiners Anna Alexandrova and Anna Mahtani for taking time to read through this thesis despite the additional stress caused by the corona pandemic.

Finally, I want to thank Bjarke Mønsted for our joint collaboration and his enormous support during the final stages of my PhD. I have hugely enjoyed working with you - both theoretically and practically - and I look forward to many future projects together ♡.

Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

The aim of this thesis is to argue that one way in which dissemination of scientific descriptions can have unintended social implications is by promoting social norms that influence people's behaviour in certain situations. In order to make this argument, I use the case of microeconomics and argue that microeconomics textbooks and teaching practices can make people follow a social norm of self-interest - and often narrow self-interest - in economic situations. The argument thus presents one possible way microeconomics can be self-fulfilling and have unintended social implications. This fills a gap in the philosophy of science literature since no one - to my knowledge - has made a satisfactory argument for how microeconomic descriptions come to influence behaviour. It also contributes to the field of social ontology by showing that two social norms (one of cooperation and one of self-interest) exist in economic situations. Finally, the analyses in the thesis show that even if it is possible in principle to distinguish between positive and normative microeconomics, this distinction does not hold historically nor in contemporary microeconomics textbooks. This contributes to the methodological concerns in philosophy of economics since it questions whether it is desirable or even attainable to maintain the distinction.

Microeconomics is a branch of economics which - traditionally - is concerned with the decision making of and interaction between agents when

allocating scarce resources.[1] Looking at the current scope of microeconomics textbooks, the situations studied in microeconomics extend to situations concerning consumers and consumption, choices under uncertainty, coordination and competition with other agents, and bargaining and market situations. Throughout the thesis, I will refer to these situations as *economic situations*.[2]

Orthodox microeconomics textbooks use mathematical models to study economic situations under idealised circumstances.[3] Though the models depict different situations, they all share the assumption that agents act rationally [Morgan, 2012, p.394]. This assumption comes in different variants depending on the model, but common for all variants is that they assume agents to have rational preferences such that their preference relations are complete and transitive. Further, they assume that rational agents only care about their *individual gains*. Here, I use *gains* as a place holder that can refer to monetary gains, preferences, or utilities. Finally, all variants of the assumption state that rational agents strive to *optimize* their choices with regard to their possible gains. For short, I will refer to all variants of this assumption considered in this thesis as *the rationality assumption*. I will go through a detailed description of how these variants of the rationality assumption differ in section 5.3. For now, it suffices to note that in models where gains are interpreted as preferences or utilities it is sometimes possible for agents to have other-regarding preferences. Thus, in some microeconomic models, the rationality assumption need not imply that agents act self-interestedly.

Despite this fact, several variants of the rationality assumption do imply that agents act self-interestedly. Further - as I will argue in chapters 5 and 6 -

[1] Here the term *agent* is understood broadly so as to include both individuals and firms. See List and Pettit [2011, ch.1] for an account of firms as agents.

[2] See subsection 6.3.2.

[3] Microeconomics textbooks used at university level differ surprisingly little. In the thesis, I have chosen to focus on a selection of textbooks used at the top economics departments. These are Mas-Colell et al. [1995], Jehle and Reny [2011], and Varian [2014]. Chapter 5 provides a further discussion of and introduction to the textbooks. Since the scope of the thesis is orthodox microeconomics as it is taught at university level, I will not discuss how or whether other branches of microeconomics (such as behavioural economics or heterodox microeconomics) is taught. Further, I will only discuss microeconomics papers that are used in traditional microeconomics teaching.

microeconomics textbooks present and discuss all variants of the rationality assumption in the standard models as if they imply that individuals act self-interestedly. This focus on self-interest in microeconomics textbooks is made clear in the first sentence of the widely used textbook *Microeconomic Theory* by Mas-Colell, Whinston, and Green [1995]:

> A distinctive feature of microeconomic theory is that it aims to model economic activity as an interaction of individual economic agents pursuing their private interests [Mas-Colell et al., 1995, p.3].

Here, I use *self-interest* to describe situations where agents only care about their self-regarding preferences and utilities. I use *narrow self-interest* to describe situations where agents only care about their own monetary gains [Bicchieri, 2006, p.105]. Finally, I use *(narrow) self-interest* as an abbreviation when I speak about both types of behaviour.[4]

Looking at how scientific descriptions can influence the world, I use the concept *self-fulfilling science* to denote a science that includes descriptions of how all agents act and where the dissemination of these descriptions has social implications by influencing people's behaviour to be more like how it is described.[5] The influence can be understood as unintended, if the descriptions are not aimed at changing people's behaviour but rather at describing it or predicting it (by using false assumptions). Given this definition, microeconomics is self-fulfilling if the rationality assumption affects people's behaviour such that they start acting more in accordance

[4] A critical reader may argue that microeconomics textbooks use "self-interest" as including other-regarding preferences, such that when economists speak of self-interest it has a different meaning than the one normally attached to it (see e.g. Bicchieri [2006, p.17]). This however, is contrary to the historical tradition in economics which describes self-interest (as normally understood) as beneficial (see chapter 4). Further, none of the textbooks analysed state that their use of self-interest differs from how it is ordinarily used. Thus, even if the authors used the word in a different sense, the readers of the textbooks will not be aware of this. Throughout the thesis, I will therefore assume that self-interested preferences are self-regarding and do not include other-regarding preferences. If one does not accept this assumption, then the scope of the thesis will be reduced to showing that a social norm of narrow self-interest is being promoted in economic situations where a monetary gain is possible.

[5] See chapter 2.

with the behavioural rules it endorses. By looking at the standard models in microeconomics textbooks - where the rationality assumption is always described as implying a behavioural rule of (narrow) self-interest - I show one way in which microeconomics can become self-fulfilling.[6]

In order to do so, it is first important to determine whether microeconomics actually is self-fulfilling. This can, for example, be done by using a subfield of microeconomics called *game theory*. Game theory studies how rational agents choose when the outcomes of their actions depend on what other agents do. The situations studied in game theory are called games. By examining how real people play these games, it is possible to determine whether people's behaviour conforms to the theoretical solutions used in game theory. It can also tell us whether people's behaviour changes when they are exposed to different stimuli. Finally, it makes it possible to test whether economists choose more in accordance with the rationality assumption compared to non-economists. Several studies have tested this latter point, and I provide a comprehensive review of their findings in chapter 3. In the remainder of chapter 1, I first define some basic concepts and games used in game theory. Next, I motivate the argument of the thesis by presenting the results of the first experiment conducted to test whether economists and non-economists act differently in economic situations. I end the chapter by sketching the main argument and structure of the thesis.

[6]Since the focus of the thesis is on microeconomics as it is presented in standard models in microeconomics textbooks, and since textbooks do not account for the fact that some of these variants of the rationality assumption need not imply self-interested behaviour, it may be questioned whether the scientific descriptions in microeconomics textbooks actually count as microeconomic descriptions at all. Though it is important to acknowledge the difference between a scientific theory per se and the communication of and education in that theory, I will none the less maintain that what is taught in microeconomics courses at university level (to bachelor's, master's, and PhD students) *is* microeconomic theory. I do this, since microeconomic theory would otherwise only be known to a selected handful of academics. Further, if advanced microeconomics textbooks do not provide real microeconomic descriptions, then it is questionable what will. Thus, the thesis builds on the assumption that microeconomics textbooks do contain scientific descriptions and that if they can be shown to promote a social norm of (narrow) self-interest, then this will show one way that microeconomics is self-fulfilling.

Some game theoretical concepts

Throughout the thesis, I will consider two types of games when examining whether microeconomics can have an effect on individuals' behaviour. The first type of game is a *normal form game* or strategic form game. A normal form game is specified by 1) the players in the game, 2) the strategies - i.e. a set of moves - available to each player in the game, and 3) the payoffs each player will get depending on which strategies all players in the game choose [Gibbons et al., 1992, pp.115-116]. In normal form games, players will typically choose simultaneously without knowledge of each other's choices. The second type of game is an *extensive form game* [Kuhn and Tucker, 1953]. An extensive form game is specified by 1) the players in the game, 2) elements of chance or "nature" in the game, 3) the strategies available for each player in the game including when the player can make a move, and what information each player has per move they make, and 4) the payoffs of each player for all possible strategy combinations [Jehle and Reny, 2011, pp.325-327]. Extensive form games can - for example - be used to study game situations where the players make choices sequentially. The payoffs are given by real numbers that usually represent either utilities or monetary payoffs.

Games can be represented in different ways. All games can be represented mathematically by a tuple specifying the elements of the game described above. However, simple two player normal form games will typically be represented by a matrix - called a *matrix representation* - while simple two player extensive form games typically will be represented by a decision tree - called a *diagrammatic representation*. Notice, however, that both simple normal form games and simple extensive form games can be represented using a matrix representation or a diagrammatic representation.[7] The fact that simple normal form games are typically represented in a matrix while simple extensive form games are typically represented diagrammatically is

[7] Originally, Von Neumann and Morgenstern [1944b] defined normal form games as games represented by a matrix and extensive form games as games represented by a decision tree. However, since it is possible to represent a normal form game by a decision tree and an extensive form game by a matrix, I have chosen to use the current textbook terminology, where a clear distinction is made between the game form and the representation of the game.

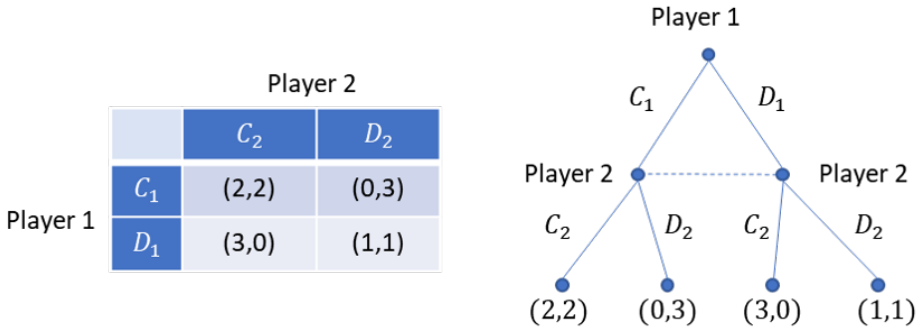


Figure 1.1: **Illustration of prisoner's dilemma.** The game is illustrated in a matrix representation (L) and diagrammatic representation (R).

primarily a matter of convenience and tradition.

Figure 1.1 shows the matrix representation (left) and the diagrammatic representation (right) of the same normal form game. In the matrix representation, the rows of the matrix depict the two available strategies for player 1 (C_1 and D_1) while the columns depict the available strategies for player 2 (C_2 and D_2). Each matrix entry specifies the payoffs for the two players if the corresponding strategy combination is chosen. The first number indicates the payoff of player 1 while the second number indicates the payoff of player 2. In the diagrammatic representation, the two available strategies for player 1 are depicted in the top. The top node in the tree is thus a decision node for player 1, where the two edges show the two possible moves for player 1.[8] The two middle nodes are decision nodes for player 2. Notice that player 2 have the same options independently of what player 1 chooses. Further, the dotted line between player 2's two decision nodes indicates that player 2 does not know which node they are at.[9] In other words, the dotted line indicates the information set for player 2. Finally, the bottom nodes are called terminal nodes. Here, the payoffs for following each specific path through the tree are depicted. If the payoffs are depicted horizontally (as in figure 1.1), the payoff to the left is the payoff of player 1. If the payoffs are depicted

[8]Since player 1 does not have any further moves available in this game, the two edges are also equal to player 1's two strategies.

[9]I will employ the singular "they" throughout the thesis when providing examples where the gender of an agent is irrelevant.

vertically, the upper payoff is the payoff of player 1.

The game depicted in figure 1.1 is a version of a famous game called the *prisoner's dilemma* (PD) [Flood et al., 1950]. The game is used extensively in game theory and in the experimental literature on economists' behaviour. We will therefore revisit it several times throughout the thesis. In the story accompanying prisoner's dilemma, two prisoners have to independently decide whether to *cooperate* in order to get a reduced time in jail, or to *defect* from cooperation. The dilemma arises because a police officer has told each prisoner separately that if they defect (and the other prisoner does not), then they will go free while the other prisoner will get a long jail sentence (e.g. five years). However, if both prisoners defect, they will both get longer time in jail (e.g. three years) compared to if they both cooperate (e.g. one year). Assuming that reduced time in jail means higher utility, figure 1.1 depicts just this situation, where the payoffs are interpreted as ordinal utilities. Here then, C_1 is interpreted as player 1 cooperating, while C_2 is player 2 cooperating. Likewise, D_1 means that player 1 defects while D_2 means that player 2 defects.[10]

The prisoner's dilemma is famous because its solution strikes people who have not studied game theory as counter intuitive. To see why, we first have to introduce the game theoretical solution concept of a *strictly dominant strategy*. A player's strategy is strictly dominant if it will lead to the highest possible payout for the player independently of what other players choose. If both players in the prisoner's dilemma are rational, they will both seek to maximise their payoffs. Since neither player knows what the other player will choose, each player will have to consider all possible strategy combinations in order to make the best choice. Consider the choice of player 1: player 1 knows that player 2 can either cooperate or defect. If player 2 cooperates, player 1 will gain the highest payoff by defecting, since $3 > 2$. If player 2 defects, player 1 will still gain the highest payoff by defecting, since $1 > 0$. Thus, defection is a strictly dominant strategy for player 1, since it will maximise their payoff irrespectively of what player 2 chooses. Since the

[10] A more precise definition prisoner's dilemma is presented in chapter 3, section 3.2.

game is symmetric, player 2 will also maximise their payoff by defecting.[11] Thus, the game theoretical solution to the prisoner's dilemma is that both players will defect. This is surprising, because both players would have gained more if they had both cooperated.

The prisoner's dilemma is a very simple game that shows how acting in accordance with the rationality assumption can lead to an outcome where everyone is worse off than they could have been. Such an outcome - where it is possible to make at least one player better off without making anyone else worse off - is called a *Pareto inferior outcome*. An outcome, where it is impossible to make anyone better off without making someone else worse off, is called a *Pareto optimal outcome*. Since there is a long-standing tradition in microeconomics claiming that it is publicly and individually beneficial to act in accordance with one's own interests (see chapter 4), it is a surprising finding in game theory that rational behaviour does not always lead to a Pareto optimal outcome.

Another game with a similarly surprising outcome is the public goods game (PGG). A standard version of the PGG is played with four players.[12] Each player is given a number of tokens, for example 10, which they can invest in an individual exchange or in a public exchange. The tokens they invest in the individual exchange will be translated into a monetary amount that they will receive with certainty, e.g. 50 pence per token. The tokens they invest in the public exchange will be translated into money at a higher exchange rate, for example £1 per token. However, the money gained from the public exchange will be shared equally between all four players regardless of who invested the tokens. This means that if all players have 10 tokens and only invest in the individual exchange, each player will receive $10 \cdot £0.5 = £5$. Further, if all players invest their 10 tokens in the public exchange, each player will receive $\frac{10 \cdot 4 \cdot £1}{4} = £10$. Collectively, then, everyone is better off if all players invest all their tokens in the public exchange.

Consider now a case where three players invest all their tokens in the

[11] In a symmetric game, all players have the same set of strategies and the payoff structure for the strategies is the same for all players [Jehle and Reny, 2011, p.365].

[12] See for example Keser [1996], Croson and Marks [1998], Willinger and Ziegelmeyer [1999], and Cookson [2000].

public exchange while the fourth player invests their tokens in the individual exchange. The first three players will then receive $\frac{10 \cdot 3 \cdot £1}{4} = £7.5$ while the fourth player will receive $£7.5 + 10 \cdot £0.5 = £12.5$. Further, if three players invest their tokens in the individual exchange and one player invests their token in the public exchange, then the one player will receive $\frac{10 \cdot £1}{4} = £2.5$ while the three players will each receive $£2.5 + 10 \cdot £0.5 = £7.5$. Assuming that player 1 invests t_1 of their token in the public exchange, we can write the total monetary return, MR_1, for player 1 as

$$\begin{aligned} MR_1 &= £0.5 \cdot (10 - t_1) + \frac{£1 \cdot (t_1 + t_2 + t_3 + t_4)}{4} \\ &= £0.5(10 - t_1) + £0.25 \cdot t_1 + £0.25 \cdot (t_2 + t_3 + t_4). \end{aligned}$$

Since player 1 cannot influence how much the other players invest in the public exchange, $£0.25 \cdot (t_2 + t_3 + t_4)$ can be seen as a exogenous variable. This leaves player 1 with the decision of whether to invest in the individual exchange - with an exchange rate of £0.5 - or to invest in the public exchange with an exchange rate of £0.25. Thus, the strictly dominant strategy for player 1 is to invest all their tokens in the individual exchange because $£0.5 \cdot t_1 > £0.25 \cdot t_1$ for $t_1 > 0$. Since the game is symmetric, the game theoretical solution to the PGG is that no rational player will invest in the public exchange.

The experiment that started it all

In 1981, Marwell and Ames published the results from 12 experiments conducted to see what people actually choose when faced with a PGG under varying circumstances.[13] They used a PGG like the one described above, but where all participants invested in the same public exchange and where the public exchange rate was an increasing function of the number of tokens invested in it (starting at a lower rate than the individual exchange rate). Table 1.2 summarises the findings of the 12 experiments.[14]

Marwell and Ames drew several conclusions from the experiments. First, the game theoretical solution to a PGG - that participants will only invest

[13] There were 32 participants in each experiment. However, participants in the experiments were told that they were in groups of 80 people. The participants were high school students

	Experiment	Average investment
1.	Basic PGG.	42%
2.	Difference in initial number of tokens and/or in public exchange rates.	53%
3.	Minimum investment requirement for the public exchange.	51%
4.	PGG in groups of 4.	60%
5.	Experienced participants.	47%
6.	Higher exchange rates (factor 5).	28%
7.	Two investment rounds.	46%
8.	Two investment rounds. Possibility to reallocate tokens invested in the first round.	50%
9.	Two investment rounds. College students as participants.	49%
10.	Two investment rounds. Manipulated feedback in how much is invested in the public exchange after the first round (low, medium high).	43%, 50%, 44%
11	Public exchange as contribution to shared group project.	84%
12	Economics students and higher returns (factor 2).	20%

Table 1.2: **Summary of Marwell's and Ames's 12 experiments.** Left column: the 12 experiments. Right column: the average investment in the public exchange per experiment in percentage of tokens invested.

in the individual exchange - is contradicted by evidence since participants usually invested between 40% and 60% in the public exchange [Marwell and Ames, 1981, p.307]. Second - and providing a point of departure for this thesis - the economics students participating in experiment 12 behaved differently from the other participants in the experiments (who had not

except in experiments 9 and 12 where the participants were college students.

[14]In experiment 6, Marwell and Ames employed a new, inexperienced interviewer. The participants interviewed by the new interviewer contributed less to the public exchange compared to the other participants. The result of experiment 6 excluding the participants interviewed by the new interviewer is 35%.

received training in economics). Indeed, Marwell and Ames [1981, p.306] start the result subsection for experiment 12 with the statement "At last, a result that is really different". The experiment showed that economics students are less likely to invest in the public exchange compared to any other group of participants. Thus, they acted more in accordance with the rationality assumption compared to other participants. This provides the first evidence supporting the argument that microeconomics is self-fulfilling.

The finding was further strengthened by Marwell's and Ames's research on participants' conception of fairness in the PGG [Marwell and Ames, 1981, pp.308-310]. After the experiments had ended, each participant was asked what they thought a fair investment in the public exchange would be, and how concerned they had been with fairness when choosing how much to invest in the public exchange. Participants from experiments 1 to 11 generally answered that it is fair to invest 50% or more in the public exchange. Further, there was a positive correlation between participants who invested a lot in the public exchange and participants who had indicated i) that they were concerned about fairness when making their decision and ii) that it would be fair to invest a high number of tokens in the public exchange. Comparing these answers to the answers from the economics students in experiment 12, Marwell and Ames [1981, p.309] write:

> Comparisons with the economics graduate students is very difficult. More than one-third of the economists either refused to answer the question regarding what is fair, or gave very complex, uncodable responses. It seems that the meaning of 'fairness' in this context was somewhat alien for this group. Those who did respond were much more likely to say that little or no contribution was 'fair'. In addition, the economics graduate students were about half as likely as other subjects to indicate that they were 'concerned with fairness' in making their investment decision.

Thus, Marwell and Ames suggest that one reason economics students acted differently than other participants in the experiment is that they have a different conception of fairness and may not care as much about fairness in this situation as other people do.

The paper initiated an ongoing debate considering whether economists and economics students act differently than non-economists. Specifically, the debate has sought to answer whether economists act more in accordance with the rationality assumption compared to others and why this is. Using the terminology of this thesis, the debate initiated by Marwell and Ames [1981] sought to determine whether microeconomics is self-fulfilling, and - if yes - why it is so. Thus, the thesis contributes to this debate by arguing, first, that the current empirical literature supports the conclusion that microeconomics is self-fulfilling, and, second, that one way microeconomics can be self-fulfilling is by promoting a social norm of (narrow) self-interest in economics situations (like the PGG).

The main argument

The claim that exposure to microeconomics textbooks and teaching practices can promote a social norm of (narrow) self-interest may seem puzzling, since Friedman [1953] famously argued that we can distinguish between positive and normative economics, and that positive economics (aiming at predicting behaviour) is independent of normative judgements. Friedman's argument has been very influential in contemporary economics, and his paper is one of the few papers that economics students are typically asked to read [Mäki, 2009]. Before presenting the argument for how microeconomics can be self-fulfilling, I therefore turn to a historical analysis of economic theory in order to refute Friedman's claim that positive microeconomics is free from normative judgements.[15] The analysis shows that modern economics since its beginning has been influenced by the 18th century idea that it is beneficial to control people's wilder passions of power and lust with the calmer passion of greed. Further, the claim that (narrow) self-interest is publicly and individually beneficial is still present in contemporary positive economics. This suggests that Friedman's distinction between positive and normative economics may not be as easily attained or desirable as it is claimed to be.

[15] See chapter 4.

The analysis also shows that the arguments stating that (narrow) self-interest is beneficial are easily refuted. This raises the questions why the behavioural rule of (narrow) self-interest is still implied by the textbook descriptions of the rationality assumption and how the current textbooks describe the assumption. In chapter 5, I answer these questions, first, by analysing the historical development of the rationality assumption and, second, by presenting the different variants of the assumption as they are used and explained in the standard models in contemporary microeconomics textbooks. By looking at the development of the rationality assumption, we see that it is such a prominent assumption in microeconomics because it is closely related to the development of economics as a separate scientific discipline employing mathematical models. This fact also helps explaining why the assumption is used in all microeconomic models. Presenting the current variants of the rationality assumption used in the standard positive models in microeconomics textbooks, I finally show how the assumption is both defended as descriptively plausible and normatively desirable. Thus, even though the textbooks distinguish between positive and normative economics, they fail to keep their account of positive microeconomics free from normative judgements.

Equipped with these analyses, I turn to the question of how microeconomics textbooks and teaching practices at universities can promote a social norm of (narrow) self-interest among the students in economic situations. According to Marwell and Ames [1981], economics students acted differently in the PGG because they had a different conception of fairness compared to other participants. In a recent paper, however, Gerlach [2017] argues that economics students act differently in economic situations because they *expect* other people to behave in that way:[16]

> This study demonstrated that, relative to their fellow students, economics students [...] were about equally likely concerned with fairness, and they had a similar understanding of what was fair. However, economics students expected to receive smaller

[16] Participants in Gerlach [2017] were asked to play a dictator game. For further information, see section 3.2.

> offers from others, which in turn mediated their own smaller offers. Moreover, economics students were less willing to veto unfair allocation of others. Taken together, the results suggest that economics students' more selfish behavior is not due [to] different fairness standards but to social norms [Gerlach, 2017, p.10].

There are several different approaches to defining and studying social norms [Bicchieri et al., 2018]. In this thesis, I use Bicchieri's [2006] definition of social norms, as presented in *The Grammar of Society: The Nature and Dynamics of Social Norms*.[17] Here, Bicchieri [2006, p.11] provides the following conditions for a social norm to exist and be followed:

> **Definition 1** *Conditions for a Social Norm to Exist:*
> Let R be a *behavioral rule* for situations of type S, where S can be represented as a mixed-motive game. We say that R is a social norm in a population P if there exists a sufficiently large subset $P_{cf} \subseteq P$ such that, for each individual $i \in P_{cf}$:
>
> ***Contingency:*** i knows that a rule R exists and applies to situations of type S;
>
> ***Conditional preference:*** i prefers to conform to R in situations of type S on the condition that:
>
> **(a)** ***Empirical expectations:*** i believes that a sufficiently large subset of P conforms to R in situations of type S;
>
> and either
>
> **(b)** ***Normative expectations:*** i believes that a sufficiently large subset of P expects i to conform to R in situations of type S;
>
> or
>
> **(b')** ***Normative expectations with sanctions:*** i believes that a sufficiently large subset of P expects i to conform to R in situations of type S, prefers i to conform, and may sanction behavior.

[17] I have chosen to use this definition because it is operational such that it can be tested empirically whether people satisfy its conditions. Further, the definition is accepted and used in a wide variety of scientific disciplines, including philosophy and the experimental literature referred to above. By using the same definition, I thus increase the consistency and comparability between the relevant literature and this thesis. For further discussion of the definition, see section 6.2.

> A social norm R is *followed* by a population P if there exists a sufficiently large subset $P_f \subseteq P_{cf}$ such that, for each individual $i \in P_f$, conditions 2(a) and either 2(b) or 2(b') are met for i and, as a result, i prefers to conform to R in situations of type S.

In the definition, Bicchieri distinguishes between the existence of a social norm and a social norm being followed. Throughout the thesis, I will use *inclined to follow a social norm* about norm followers who know that a social norm exists and who follow that norm unless they receive sufficient information to conclude that conditions 2(a) and 2(b) are not satisfied in a situation. Further, I say that a social norm is *stabilised* if sufficiently many people will continue to satisfy conditions 2(a) and 2(b) in a situation so that they will keep following the social norm. Finally, I will say that something *promotes* a social norm, if it makes people inclined to follow the norm and if it helps stabilising the norm such that people will keep following it.

Using the definition of social norms as a starting point, I argue that readers of microeconomics textbooks will be inclined to follow a social norm of (narrow) self-interest in economic situations, since microeconomics textbooks inform people that the behavioural rule exists in these situations (condition 1). Further, I use the analyses of how the textbooks describe and defend the rationality assumption to argue that they can make the readers believe that people in actual economic situations follow the behavioural rule of (narrow) self-interest (condition 2(a)) and that people will expect them to follow the same rule (condition 2(b)). Given this, microeconomics textbooks can make their readers inclined to follow the social norm of (narrow) self-interest in economic situations.

Considering the structure and content of microeconomics courses at universities, I further argue that these courses will confirm the students' expectations in conditions 2(a) and 2(b) such that the social norm will be stabilised. This is the case because the teaching practices do not leave room for critical discussion of the rationality assumption. Instead, the lectures, class teaching, and assignments focus on how to manipulate microeconomic models (using the rationality assumption) and how to apply these models to different economic situations [Earle et al., 2016]. Thus, the students'

expectations will be confirmed by their teachers and by their peers who are trying to understand the same microeconomic models. Finally, I argue that the students can form a hub in their social network from where the social norm of (narrow) self-interest can spread and be adopted by people who have not themselves been exposed to economic theory. Thus, microeconomics textbooks and teaching practices may have social implications that affect people's behaviour beyond the classroom.

The argument sketched above presents one possible way that microeconomics can become self-fulfilling. However, the argument does not show whether the changes in behaviour observed for economics students are actually caused by a social norm.[18] In order to show this, we conducted three experiments designed to test whether the above argument is supported empirically [Buchter et al., 2020].[19] The first experiment tests whether we can influence people's choices in a prisoner's dilemma game by exposing them to microeconomic terminology. The results show that exposure to microeconomic terminology makes people defect more than people who have not been exposed to microeconomic terminology. They also show that economics students defect more than other participants regardless of which terminology they are exposed to. In the second experiment, we tested whether the observed changes in behaviour are caused by a social norm. The experiment confirms this and shows that in economic situations - like the prisoner's dilemma - there exist a social norm of cooperation and a social norm of (narrow) self-interest. Finally, the experiment shows that exposure to microeconomic terminology changes people's behaviour by making them satisfy conditions 2(a) and 2(b) with regard to the social norm of (narrow) self-interest. In the third experiment we conducted several simulations based on the findings from the first experiment to see how exposure to microeconomic terminology can influence behaviour in a social network constructed from data on the interactions between 1000 university students [Stopczynski et al., 2014]. The results from the experiment show that under certain circumstances,

[18]Throughout the thesis, I will use an interventionist definition of causation. See for example Woodward [2005] and Woodward [2010].

[19]See appendix to this thesis.

the behavioural effects caused by exposure to microeconomic terminology can push the entire network over a tipping point, such that everyone in the network will start defecting. Thus, the results support the argument that microeconomics textbooks and teaching practices can influence people's behaviour and have social implications beyond the classroom.

The structure of the thesis

The thesis is divided into two parts. In part I, I present the philosophical and historical background analyses needed to make the main argument which I then present in part II.

Part I consists of four chapters. In chapter 2, I discuss the claim that science - and especially social and human sciences - can change people's behaviour by describing it. I argue that there are subtle differences between the concepts used to describe this phenomenon and between the meanings different theorists attach to the same concepts. To remedy this, I propose to consider the phenomenon along seven dimensions where the concepts differ. Further, I argue that the differences between the meanings attached to the concepts can be explained by the focus that each theorist has when considering the phenomenon. Thus, I propose that we order the future debate by distinguishing between different relevant issues related to the phenomenon. Finally, I propose to use the concept *self-fulfilling science* when describing how the dissemination of scientific descriptions aimed at all agents' behaviour can have social implications by changing people's behaviour. In chapter 2, I thus set the stage of the thesis by exploring the key theoretical ideas from philosophy of science that has motivated it.

In chapter 3, I review the current empirical literature comparing the behaviour of economists and non-economists. Based on the empirical results, I argue that microeconomics is a self-fulfilling science and that this effect can occur via exposure to the rationality assumption as it is presented in microeconomics textbooks.

In chapters 4 and 5 I provide a historical analysis of microeconomic theory. In chapter 4, I argue that positive economics has been closely connected to the normative claim that it is beneficial to act (narrowly) self-interested

since the beginning of classical economics. Further, I show how these normative claims are still present in contemporary neoclassical economics. This questions whether Friedman's [1953] distinction between positive and normative economics is viable.

In chapter 5, I argue that one reason the rationality assumption is so strong and persistent in microeconomics is that its development is closely related to the development of economic theory. Further, I present the different variants of the rationality assumption as they are described in the standard models by contemporary microeconomics textbooks and argue that the textbooks do not satisfy the distinction between positive and normative economics since they defend the rationality assumption used in positive economics as both empirically plausible and normatively desirable.

In part II, I use the analyses of microeconomic theory to argue how microeconomics textbooks and teaching practices can promote a social norm of (narrow) self-interest. In chapter 6, I present a theoretical argument for how this can occur by referring to the psychological mechanisms used by Bicchieri to explain how social norms emerge and change. In chapter 7, I present the results from three empirical studies we conducted to test the plausibility of the argument. Taken together, the two chapters provide a detailed account of how microeconomics can become self-fulfilling by promoting a social norm of (narrow) self-interest in economic situations via the dissemination of microeconomics textbooks and teaching practices. I conclude the thesis in chapter 8, where I also provide an outlook and discuss two possible ways to mitigate the social implications of microeconomics textbooks and teaching practices.

Part I

PHILOSOPHICAL AND HISTORICAL BACKGROUND

CHAPTER 2

SELF-FULFILLING SCIENCE

2.1 Introduction

Consider the lattice structure of a salt crystal. It is a common assumption that this structure will not be affected by how our scientific theories describe it. As a general statement, we might say that knowledge of scientific descriptions cannot by itself change the world. In the past century, however, this general claim has been disputed by some sociologists and philosophers. Instead, they argue that - especially in the human and social sciences - merely describing the world can change it, making it more or less similar to the description.

If knowledge of scientific descriptions by itself can change the world, then this phenomenon will impact several different problems discussed in philosophy of science. First, assuming that the phenomenon only occurs in the human and social sciences, one may wonder whether these sciences can be held to the same methodological standards as the natural sciences. Second, the phenomenon can impact how we view the function of science in society and our use of scientific theories in the world. Finally, the phenomenon may have social implications for our society that ought to be reflected in the way we do and use science. For all three problems, it is relevant to ask whether the phenomenon does in fact occur, and - if yes - which mechanisms can

cause it.

The project of the thesis focuses on the social implications of the phenomenon and answers the two questions of whether it occurs and - if yes - how it can occur. I do this by using the case of microeconomic theory as it is currently taught at universities and argue that one in which microeconomics can become self-fulfilling is via the promotion of social norms. By answering these questions, I contribute to the debates in philosophy of science and social ontology by showing how scientific descriptions can have social implications and that people's actions in economic situations are guided by social norms.

Before answering these questions, however, it is worthwhile to explore what has already been said about the phenomenon. I therefore start this chapter by discussing the different theories in philosophy of science that examine the phenomenon. These are the theories of self-fulfilling prophecies and reflexive predictions, looping effects, and performativity. Although many theorists treat the concepts as largely synonymous, there are subtle differences between them and between the meanings different theorists attach to the same concept. These differences call for clarification and analysis. In the second part of the chapter, I therefore provide a taxonomy of seven relevant dimensions for the phenomenon. I further propose to redefine the concepts used to describe the phenomenon, so that each concept relates to one of the three problems stated above. Redefining the concepts in this way will prevent the current confusion in the debate, where it is not always clear how a concept is defined or which problem theorists are trying to address.

The aim of this chapter is thus to set the stage for the thesis by exploring the key theoretical ideas and motivation underlying it. In doing this, I also clarify the differences among and internal to the theories examining whether knowledge of scientific descriptions can change the world, and provide a taxonomy to discuss the relevant dimensions of the phenomenon.

2.2 Changing the world by describing it: discussion of the different theories

Theories concerning the phenomenon that describing the world can change it, can be sorted into three groups. The first group consists of theories primarily using the concepts of *self-fulfilling prophecies* and *reflexive predictions*. To account for these concepts, I discuss the arguments by Merton, Popper, Nagel, Romanos, Buck, and Kopec. The second group uses the concept of *looping effects* which was introduced by Hacking. Finally, the third group of theories uses the concept of *performativity* and can - for example - be found in the writings of Callon and MacKenzie. In order to provide a thorough account of the phenomenon, I will look at each of the three groups in turn, highlighting the differences and similarities between the concepts, as well as the different definitions associated with each concept.

2.2.1 Self-fulfilling prophecies and reflexive predictions

The idea behind the first group of theories can be traced back to *Child in America* by Thomas and Thomas [1928]. Here, they state that

> the subject's view of the situation, how he regards it, may be the most important element for interpretation. For his immediate behavior is closely related to his definition of the situation, which may be in terms of objective reality or in terms of subjective appreciation - "as if" it were so. Very often it is the wide discrepancy between the situation as it seems to others and the situation as it seems to the individual that brings about the overt behavior difficulty. [...] If men define situations as real, they are real in their consequences [Thomas and Thomas, 1928, p.572].

Their focus is on children's behaviour, but their point is presented more generally; that how a person perceives or interprets a situation will affect how they act in that situation. Thus, the statement *if men define situations as real, they are real in their consequences* - known as the "Thomas theorem" [Waller and Hill, 1938, Merton, 1995] - should be understood, not as a person's ability

to change a situation, but as saying that a person's understanding of the situation has a real effect on how the person decides to act in that situation.

In 1948, Merton introduced the term *self-fulfilling prophecy* in order to discuss the implications of the Thomas theorem for society [Merton, 1948, p.193]. According to Merton:

> The self-fulfilling prophecy is, in the beginning, a *false* definition of the situation evoking a new behavior which makes the originally false conception come *true* [Merton, 1948, p.195].

Here, we see that self-fulfilling prophecies differ from the Thomas theorem by having a narrower scope. Where the Thomas theorem does not specify how a person's understanding of a situation relates to the situation, self-fulfilling prophecies require that the description of the situation is initially false. Further, self-fulfilling prophecies only occur when this false description invokes a behaviour which makes the description come true. This is also in contrast with the Thomas theorem, where there is no specification of how the outcome of a person's behaviour should relate to the person's original understanding of the situation. Notice that Merton does not require the initial false description to be a prediction. Thus, *prophecy* does not refer to the type of initial description, but to the process of the description changing its truth value.

Merton uses the concept of self-fulfilling prophecies to discuss how discrimination can have a self-enforcing effect. If African Americans in the 1940s are - wrongly - described as being strike breakers (because they do not have a tradition for unions), this misconception may exclude them from unions, which in turn restricts their job options. Because of their restricted options, it will be difficult for an African American not to work during a strike, which in turn makes the initially false description true [Merton, 1948, pp.196-197]. Finally, Merton argues that self-fulfilling prophecies are restricted to human affairs since they cannot be found in the world of nature [Merton, 1948, p.195].[1]

[1] It is not obvious that the scope of Merton's definition is actually restricted to human affairs. Whether Merton's definition can be extended - for example to some parts of the animal kingdom - is outside the scope of this thesis. As we shall see in this subsection,

In *The poverty of historicism* from 1957, Popper argues that the same methodology should be applied to the natural and social sciences [Popper, 1957, p.120]. As a part of his argument, Popper discusses an idea similar to self-fulfilling prophecies which he terms the *Oedipus effect*. The Oedipus effect is defined as the phenomenon when a prediction - or a piece of information - influences the situation which it was about [Popper, 1957, p.11]. According to people who - contrary to Popper - advocate different methodologies for the natural and social sciences, the Oedipus effect is a problem for the social sciences since the predictions and the phenomena they predict are both social happenings that can interact with each other [Popper, 1957, pp.12,13]. They therefore conclude that unlike the natural sciences, the social sciences cannot always aim for objective truth:

> The interaction between the scientist's pronouncements and social life almost invariably creates situations in which we have not only to consider the truth of such pronouncements, but also their actual influence on future developments. The social scientist may be striving to find the truth; but, at the same time, he must always be exerting a definite influence upon society. The very fact that his pronouncements *do* exert an influence destroys their objectivity [Popper, 1957, pp.13-14].

Popper refutes this conclusion, first, by using Bohr's argument that observing a phenomenon can change it [Bohr, 1928, 1922, p.16] to argue that the Oedipus effect can be found in all sciences [Popper, 1957, p.12]. Second, Popper notes that engineering and natural science inventions - such as a rocket - also can have a great influence on the social world [Popper, 1957, p.144]. Thus, Popper concludes that the presence of an Oedipus effect does not show that social and natural sciences should have different methodologies. Notice, however, that Popper's second reason uses the *application* of science as an argument to refute *methodological concerns*. Here then, Popper confuses two distinct concerns that should not be conflated. I will return to the differences between the two concerns in section 2.3.

determining the scope of self-fulfilling prophecies is an important question for later theorists.

When defining the Oedipus effect, Popper states that the influence a prediction has on the predicted can make it happen, prevent it from happening, or something in between. Thus, the Oedipus effect differs from Merton's self-fulfilling prophecies in that it includes predictions that are true but become false, and predictions where the truth value may not change. The domain of the two concepts also differs since the Oedipus effect is defined for all sciences whereas Merton's idea of self-fulfilling prophecies is restricted to human affairs.

The phenomenon of self-fulfilling prophecies is also discussed by Nagel [1961, ch.13] in *The structure of science* when considering some methodological problems for the social sciences. To account for the phenomenon, Nagel distinguishes between *suicidal predictions* and *self-fulfilling predictions*. Suicidal predictions are information or predictions about an event that are likely to be confirmed, but - because they become public knowledge - make people change their behaviour such that the predictions end up being falsified. As an example, Nagel recounts a predicted business recession for the American economy in 1947 which did not occur because businessmen - in response to the prediction - lowered the prices of some of their goods to increase demand and thus prevent the recession [Nagel, 1961, pp.468-469]. Thus, a conscious change in behaviour made a prediction which would otherwise have been true turn out false. Nagel defines self-fulfilling predictions as predictions that are false when they are stated, but nevertheless become true because of the actions taken in response to the false predictions. Here, Nagel gives the example of a bank run on the United States Bank in 1928 which lead to its bankruptcy. The run was based on the *false belief* that the bank had financial troubles, and because of the actions taken in response to this belief, the bank ended up having financial troubles.

Having defined these concepts, Nagel argues that it is possible to establish general social laws despite these concerns [Nagel, 1961, pp.470-473]. In agreement with Popper, he further argues that the phenomenon is not restricted to the social sciences, but can occur in the natural sciences as well [Nagel, 1961, pp.469-470]. In order to show this, Nagel provides an example - inspired by Grünbaum [1956] - of a purely physical mechanism (an

automated anti-aircraft gun) that can give rise to a self-fulfilling prediction. The anti-aircraft gun consists of a gun that can be turned by an adjustment device, a computer calculating the desired position of the gun, and a system transmitting the information from the computer to the adjustment device. Imagine, first, that the computer makes a correct calculation, but that the transmitting system has a defect so that the aircraft is not hit. This, according to Nagel, corresponds to a suicidal prediction [Nagel, 1961, 469]. Imagine, second, that the computer made a wrong calculation (we can assume it did not account for a sudden strong wind) so that the aircraft would not have been hit, had the transmitting system worked properly. However, because of the defect in the transmitting system, the computer's wrong calculation results in the aircraft being hit. This, according to Nagel, corresponds to a self-fulfilling prediction [Nagel, 1961, pp.496-470].

In 1963, Buck used the term *reflexive predictions* to consider the phenomenon discussed by Merton [1949]. Informally, Buck defines reflexive predictions as predictions that cause actors to change their beliefs and act on them in a way that changes the truth values of the predictions. In order to give a stringent definition of the phenomenon, Buck introduces *dissemination status* to describe whether or not a prediction is published or revealed to certain *social actors* who are able to act on their beliefs caused by the prediction [Buck, 1963, p.360]. Given this, Buck provides the following definition:

> A prediction is reflexive if and only if:
>
> (1) Its truth-value would have been different had its dissemination status been different,
>
> (2) The dissemination status it actually had was causally necessary for the social actors involved to hold the relevant and causally efficacious beliefs,
>
> (3) The prediction was, or if disseminated, would have been believed and acted upon, and finally
>
> (4) Something about the dissemination status or its causal consequences was abnormal, or at the very least unexpected by the predictor, by whoever calls it reflexive, or by those to whose attention its reflexive character is called [Buck, 1963, pp.361-362].

According to Buck, we should accept this definition of reflexive prediction because its conditions are satisfied for all standard examples in the literature on the phenomenon. Using his definition, Buck then argues that the potential methodological problems caused by reflexive predictions can be overcome by our standard scientific methods [Buck, 1963, pp.363-365]. Furthermore, Buck argues that the example used by Grünbaum [1956] and Nagel fails to show that there can be reflexive prediction in the natural sciences, since the anti-aircraft gun cannot be said to act on beliefs. Thus, Buck [1963, pp.366-368] tentatively agrees with Merton's [1948] restriction of reflexive predictions to human affairs - at least to the extent that they necessarily involve systems that have beliefs they can act on.[2] Using the language of contemporary philosophy, we may say that on Buck's account, reflexive predictions are restricted to cases of intentional agency (whether human or otherwise).[3]

The debate in the 1960s and 1970s adopted Buck's term *reflexive predictions* as the concept describing the phenomenon under consideration [Lowe, 2018, p.348]. In this debate, Romanos [1973] provides the final definition of reflexivity. First, Romanos [1973, pp.103-104] points out that Buck's condition that reflexive predictions necessarily include systems that can act on beliefs is unwarranted, since its only defence is that this is the case in all standard examples. Instead, Romanos proposes the following definition of the necessary and sufficient conditions for a reflexive prediction:

> The formulation/dissemination style of the prediction must be a causal factor relative to the prediction's coming out true or false [Romanos, 1973, p.106].

Here, Romanos substitutes Buck's requirement of dissemination (the prediction is public) with a formulation/dissemination style. This includes

[2]Notice that this requirement means that it is possible to have a situation where the dissemination of a prediction causes a change in the truth value of the prediction, but where the prediction is not reflexive, since the change in truth value is not caused by a change in beliefs that has been acted on.

[3]See for example List and Pettit [2011, pp.19-20]. According to List and Pettit [2011, p.20], a system can be defined as an intentional agent iff it has three features: representational states, motivational states, and a capacity to process these states and act on their basis.

predictions that are disseminated as defined by Buck as well as predictions that are merely *formulated*. The formulation of predictions can be propositions which can (or could, if only formulated in a mind) convey the prediction. However, Romanos does not restrict formulations to natural languages. Rather, he says that anything - including electrical impulses, bodily movements, and puffs of smoke - which may be interpreted as expressing a prediction, can be viewed as a formulation of a prediction [Romanos, 1973, p.105]. By using this new definition of reflexive predictions, Romanos concludes that they are present both in the social and natural sciences, and that they do not pose a great methodological problem for science.

According to Kopec [2011, p.1249], the debate on reflexive predictions ended in the 1970s because the methodological problem considered came to be viewed as easily avoidable. Contrary to this consensus, Kopec argues that reflexive predictions do in fact pose a big methodological problem for the social sciences [Kopec, 2011, p.1258]. To show this, Kopec distinguishes between *weakly* and *strongly* reflexive predictions. *Strongly reflexive predictions* are inspired by Romanos's definition:

> A prediction is strongly reflexive if and only if the mode of dissemination is sufficient to switch the truth-value of the prediction from what it would be if not disseminated [Kopec, 2011, p.1252].

Here, Kopec [2011, p.1251] defines a *prediction* as an abstract object which is a proposition that states that another proposition will obtain. Further, he uses *modes of dissemination* to capture both Romano's formulation style (which he takes to be the way the prediction is made) and Buck's dissemination style (which he understands as the mode of reproducing and transmitting the prediction) [Kopec, 2011, p.1250].

Next, Kopec argues that strongly reflexive predictions only capture a proper subset of all reflexive predictions. This is because strongly reflexive predictions require a change in truth-value and so do not allow for truth- or false-making tendencies [Kopec, 2011, p.1253] - like a less severe business recession due to the actions taken in response to its prediction. Because of this, the definition cannot fully account for the scope of the methodological problem concerning reflexive predictions. In order to account for the full

scope of reflexive predictions, Kopec introduces *weakly reflexive predictions* as a general definition for the set of all reflexive predictions:

> A prediction is weakly reflexive if and only if the mode of dissemination is sufficient to change the probability of the predicted event occurring from what it would be if not disseminated [Kopec, 2011, p.1253].

According to Kopec, weakly reflexive predictions include predictions with truth- and false-making tendencies so that the predictions cause a change in the world, making the predicted events more probable without necessarily causing them to occur. Because of this, he argues that many theories are likely to be reflexive and therefore problematic to test. Thus, Kopec [2011, p.1258] concludes that reflexive predictions creates a methodological problem for the social sciences.

Kopec's definition of weakly reflexive predictions has been criticised by Lowe [2018, p.350] for being too narrow. According to Lowe, the definition of weakly reflexive predictions only includes predictions that change the probability of the exact events the predictions are about. This, according to Lowe, is too narrow, since predictions may cause changes to events without increasing the probability that the predicted events occur. To see the difference, recall the experiment conducted by Marwell and Ames [1981], discussed in chapter 1. In the experiment, Marwell and Ames consider the difference between how people act when playing a public goods game and how economic theory predicts their actions. According to game theory, people will invest nothing in the public exchange. Marwell and Ames call this the *strong free rider hypothesis* [Marwell and Ames, 1981, p.296]. They found that this hypothesis is contradicted by evidence since even economists invest something in the public exchange [Marwell and Ames, 1981, p.307]. However, Marwell and Ames also discuss a *weak free rider hypothesis*: that participants will invest less than the optimal amount in the public exchange [Marwell and Ames, 1981, p.296]. This hypothesis is supported by evidence [Marwell and Ames, 1981, p.308]. Further, the experimental results show that economists invest less in the public exchange compared to non-economists [Marwell and Ames, 1981, pp.306-307]. Thus, Marwell and Ames's findings

support the hypothesis that dissemination of economic theory changes the situation by making people invest less in the public exchange. However, this change does not make the strong free rider hypothesis - predicted by game theory - a weakly reflexive prediction, since it is possible that the probability of each person investing *nothing* will remain unchanged even if everyone starts investing *less*. If we want to include predictions that change events to be more *similar* to the predictions, but without changing the probability of the exact predicted outcomes, the scope of weakly reflexive predictions will have to be broadened.

A summary of the different concepts and how they are defined is provided in table 2.1. Here we see that the theories vary on several parameters. First, the early authors do not seem to have distinguished between subtleties that we now appreciate. This can for example be seen in the use of the words "prophecy" and "prediction" where Kopec is the first to provide an exact definition. For all authors, however, the terms are used in a broad sense that includes information, beliefs, and descriptions. For some authors - like Grünbaum, Nagel, and Romanos - predictions also include elements such as electrical impulses.

Second, considering *what* is causing a prediction, X, to change the world, all theorists agree that it is a causal change. However, we see that the earlier theorists vaguely talk about "knowledge of X" and "acting on X" while later theorists provide us with exact requirements for the dissemination of X.

Third, looking at the outcome, all theorists agree that the main concern is a potential change in the truth value of X. The focus on truth value thus seems to be one of the characteristics for this group of concepts and theories.

Finally, the main problem considered is a methodological concern for how the phenomenon affects the possibility of making predictions, testing theories, and ensuring scientific objectivity in the social sciences. Because of this focus, the debate on what domain the phenomenon is confined to (social sciences, human affairs, or all sciences) takes precedence. Notice, however, that Merton's focus differs, since he is concerned with how certain types of beliefs (like African American being strike breakers) can be harmful for society. Thus, the concepts of self-fulfilling prophecies and reflexive

Theorist	Concept	Initial situation	Change	Outcome	Domain	Main problem
Thomas and Thomas [1928]	Thomas theorem.	A believes X about Y.	A acts on X.	The action affects Y.	Human affairs.	understanding children's behaviour.
Merton [1948]	Self-fulfilling prophecy.	A falsely believes X about Y.	A acts on X.	X becomes true.	Human affairs.	Their effect on society.
Popper [1957]	Oedipus effect.	A gets information X about Y.	A acts on X in Y.	X becomes true, false, or does not change truth value.	All sciences.	Methodology and objective truth in the social sciences.
Nagel [1961]	Suicide prediction.	Prediction X about Y based on Z.	Knowledge and action on X in Y.	X is falsified.	All sciences.	Making general laws in social sciences.
Nagel [1961]	Self-fulfilling prediction.	False belief or prediction X about Y.	Acting on X in Y.	X becomes true.	All sciences.	Making general laws in social sciences.
Buck [1963]	Reflexive prediction.	Prediction X about Y.	Dissemination of X causes beliefs which are acted on in Y.	The truth-value of X is changed.	Situations involving intentional agents.	The legitimacy of testing theories in the social sciences.
Romanos [1973]	Reflexive prediction.	Prediction X about Y.	Dissemination formulation of X is a casual factor for outcome in Y.	X becomes true or false in Y.	All sciences.	Testing theories in social sciences.
Kopec [2011]	Weak reflexive prediction.	Prediction X about Y.	Dissemination mode of X casually influences Y.	The probability of X in Y is changed as a result.	Social sciences.	Testing theories in social sciences.

Table 2.1: **Summary of different concepts of reflexive predictions.** The table summarises the different concepts of self-fulfilling or reflexive predictions described above. Here, A refers to a person or a group of people, X to a prediction, Y to a situation, and Z to previous observations or data.

predictions vary in their definitions, explication of subtleties, and main areas of concern.

Do self-fulfilling prophecies apply to microeconomics?

In the remaining part of this subsection, I will consider two criticisms made by Lowe [2018] concerning this group of theories - exemplified by Kopec's two definitions. The two arguments are of special relevance for the thesis, since they consider whether self-fulfilling prophecies and reflexive predictions can be used to account for the phenomenon that people who have studied microeconomics tend to act more in accordance with its assumptions compared to people who have not studied microeconomics.

First, Lowe [2018, pp.351-352] argues that Kopec's definitions only account for *specific* predictions about *specific* situations that changes these situations. This, according to Lowe, is a problem if we consider the effect of the rationality assumption since he claims - following Ferraro et al. [2005] - that economic theory *in general* will affect people's *overall* behaviour. If a requirement for having a reflexive prediction is that it is a specific prediction about a specific situation, and the rationality assumption does not make any specific predictions, then reflexivity is unable to account for any effects caused by the rationality assumption.

According to Ferraro et al. [2005, p.17], self-interest is foundational to all economic assumptions. Indeed, they describe economics simply as stating that

a. people act self-interestedly,

b. markets are the most efficient way to organize exchanges,

c. markets are competitive, and

d. this will be beneficial for society as a whole [Ferraro et al., 2005, pp.11-12].

Thus, they present a very coarse and somewhat simplistic picture of economic theory. Further, Ferraro et al. do not define what they mean by their general

reference to "self-interested". This neglects the point (discussed in detail in section 5.3) that the rationality assumption has different variants depending on the specific situation considered. Since Ferraro et al.'s account of economics is insufficient at best, it cannot support their claim that economic theory only *in general* affects people's *overall* behaviour. Thus, Lowe's argument that reflexive predictions cannot include effects of the rationality assumption lacks evidential support.

Contrary to the claim of Ferraro et al., chapter 3 provides suggestive evidence that behavioural changes related to the dissemination of microeconomic text excerpts are connected to specific predictions about specific situations considered in microeconomic models. This is also supported by the fact - discussed in chapter 5 - that there are different variants of the rationality assumption depending on which economic situation microeconomic models are concerned with. Finally, in chapter 6, I will support this suggestive evidence with a theoretical argument stating that the behavioural effects we see are due to a social norm being promoted in specific situations corresponding to the different situations described in microeconomics textbooks.

Second, Lowe [2018, p.352] argues that the rationality assumption used in different economic theories may not be truth-apt since it is an idealisations. However, idealisations that are considered *descriptively false* - as they for example are by Friedman [1953, p.153] - are still truth-apt. Further, as I will show in chapter 5, economics textbooks do consider the different requirements concerning human behaviour in a truth-apt manner: both Mas-Colell et al. [1995, p.307] and Jehle and Reny [2011, p.267] explicitly state that they are making *descriptive* assumptions about agents' motivations and actions. Thus, this worry need not be a problem for reflexive predictions, as Lowe claims.

Since neither of Lowe's criticisms hold, self-fulfilling prophecies may be used to describe the phenomenon that knowledge of microeconomic theory can change people's behaviour. I will return to this phenomenon in section 2.3. Next, I turn to a second way the ability of scientific descriptions to cause changes in behaviour has been discussed in philosophy. This discussion was initiated by Hacking and focuses in the concepts of human kinds and looping effects.

2.2.2 Human kinds and looping effects

Hacking [1995] introduced the terms *human kinds* and *looping effects*. Human kinds are categories produced by the social sciences. They are concerned with specific groups of people, and are constructed with the aim of better understanding these groups, with the potential of intervention. Thus, human kinds are descriptions that sort people by specifying that a certain subgroup of people are *that kind* of people. Examples of human kinds are adolescent, child abuse, homosexuality, teenage pregnancy, and multiple personality. Notice that while human kinds can sort actions by sorting people who preform them, an action or specific behaviour in itself cannot be a human kind. Thus, *child abuse* as a human kind does not refer to the action of abusing a child, but to the group of people who are labelled *child abusers*, for example, because they have performed that action. Given these considerations, Hacking defines human kinds as follows:

> When I speak of human kinds, I mean (i) kinds that are relevant to some of us, (ii) kinds that primarily sort people, their actions, and behaviour, and (iii) kinds that are studied in the human and social sciences, i.e. kinds about which we hope to have knowledge. I add (iv) that kinds of people are paramount; I want to include kinds of human behaviour, action, tendency, etc. only when they are projected to form the idea of a kind of person [Hacking, 1995, p.354].

Human kinds differ from natural kinds (categorising things in the natural world) since human kinds only make sense within a certain social context. Further, human kinds are often laden with values: it is wrong to abuse children and multiple personality is a disease we want to cure. Since human kinds often convey a normative evaluation, they are something that people may want to be or may want not to be [Hacking, 1995, pp.354-355,366,367,368].[4] This leads to another difference between natural kinds and human kinds:

> If N is a natural kind and Z is N, it makes no direct difference to Z, if it is called N. However, if H is a human kind and A is a person,

[4] On pp.354-355,368, Hacking [1995] says that human kinds are *often* laden with value. On pp.366,367, he states that *all* human kinds are laden with value. On p.367, Hacking further states that *human kinds have intrinsic moral value*.

> then calling A H may make us treat A differently [Hacking, 1995, p.368].

Creating a new human kind changes how we can think of ourselves. It gives us a new vocabulary to describe experiences and thus it can change how people think of their past, how they think of themselves, and how others think of them [Hacking, 1995, pp.368-369].[5] It can also change the relationship between human kinds. It has, for example become a part of our understanding of multiple personality that it is caused by repeated childhood trauma [Hacking, 1995, p.369]. This leads us to the main difference between human kinds and natural kinds: by classifying a group of people as a specific human kind, the human kind changes the people it classifies, and so we see a looping effect.

> There is a looping or feedback effect involving the introduction of classifications of people. New sorting and theorizing induces changes in self-conception and in behaviour of the people classified. Those changes demand revisions of the classification and theories, the causal connections, and the expectations. Kinds are modified, revised classifications are formed, and the classified change again, loop upon loop [Hacking, 1995, p.370].

The greater the moral connotation of a human kind is, the more likely it is that a looping effect will occur [Hacking, 1995, p.370].[6] Looping effects can also cause the social meaning and moral value of a human kind to change, in some cases leading scientists to introduce a new word for the same phenomenon (e.g using the concept of "early parenting" instead of "teenage

[5] For further studies describing how creating a vocabulary can help people describe and voice past experiences, Hacking refers to Judith et al. [1992]. Another interesting case of this phenomenon is given by Fricker [2007, pp.149-151] when discussing hermeneutical injustice. Here, Fricker describes how the creation of the concept *sexual harassment* enabled women to talk about their present and past experiences. Further, she argues that the lack of a name for this distinctive social experience, was an epistemic injustice since it also meant an absence of understanding and collective appreciation of the phenomenon [Fricker, 2007, pp.150-151]. *Sexual harassment* can be seen as a human kind in so far as we can consider *the sexual harasser* and *the sexually harassed* as subgroups of people that we want to learn about.

[6] Unfortunately, Hacking does not provide an argument for this claim.

Theorist	Concept	Initial situation	Change	Outcome	Domain	Main problem
Hacking [1995]	Looping effect.	A is categorised as a human kind.	The categorisation changes behaviour.	Refinement of the category or new categorisation needed.	Human groups.	Their effect on society. Need of updating theories.

Table 2.2: **Summary of Hacking's concept of looping effects.** The table summarises Hacking's concept of looping effects, where A refers to a group of people.

pregnancy" or "learning disability" instead of "retarded").[7] Thus, the looping effect occurs when social scientists use value laden categorizations to consider a group of people. Knowledge of this categorization can then change how people - in and outside the group - view and behave as/towards people in the group. This change can in turn mean that the classification no longer fits the group and thus has to be changed. The theory of looping effects is summarised in table 2.2.

According to Lowe [2018], Hacking's looping effects should be seen as a group of effects of which self-fulfilling prophecies (as discussed above) is a subgroup:

> To the extent that accounts of self-fulfilling science relate to Hacking's notion, they must be viewed as an attempt to tell a kind of general story concerning one specific type of looping effect [Lowe, 2018, p.348, footnote 3].

I disagree with his assessment for three reasons. First, looping effects are intimately connected to human kinds and therefore to categorizations and not to predictions. While I do not dispute that categorization is an important part of the social sciences, I want to stress that the social sciences rely on other

[7] According to Hacking [1995, p.359], human kinds always start as scientific concepts. From there, the use can spread, and there are several examples of how people within a subgroup have tried to change the normative evaluation of the word assigned to them Hacking [1995, pp.359-360]. An example of this is the word "queer" which has been used both as a pejorative and by different sub-groups for self-identification, changing meaning through its history of use [Jagose, 1996].

reasoning styles as well.[8] As we have seen in the previous subsection, the early definitions of self-fulfilling prophecies are not connected to one specific reasoning style. If we use this broader understanding of self-fulfilment or reflexivity, looping effects can be understood as one kind of reflexivity and not vice versa. If, however, we consider the later definitions of reflexivity, we see that the concept here is constrained to predictions. While many of the categorizing effects described by looping effects may occur when scientists are making predictions about subgroups of people, it is not clear that this is the only way they can occur. Thus, using a narrow definition of reflexive predictions, the two concepts seem to describe two distinct - though closely related - phenomena.

Second, looping effects cannot be the overarching concept since it cannot account for all types of reflexive effects we see. To take the obvious example, the rationality assumption is not a human kind since it is assumed to hold for *all* agents in a model. Further, as we shall see in section 5.3, the normative defences of some of its conditions (like transitivity) state that it is beneficial for *everyone* to act in this way. Finally, the different variants of the rationality assumption do not sort people since no real person behaves exactly as they state. Thus, the assumption does not qualify as a human kind.[9] [10] If looping

[8]See for example Morgan [2012, pp.14-15] where she mentions eight reasoning styles used in science: 1) mathematical postulation and proof, 2) experiment, 3) hypothetical modelling, 4) taxonomy (classification into kinds), 5) statistical, 6) historical derivation of genetic development, 7) thinking in cases, and 8) algorithmic method.

[9]The rationality assumption is a part of some economic models that could cause a looping effect. An example of such a model is the macroeconomic *overlapping generations model* where the categorization as "old" and "young" with corresponding behavioural assumptions may have an impact on how "old" and "young" people perceive themselves and others and on how they act. Here, however, it seems to be the additional behavioural assumptions about "old" and "young" rather than the rationality assumption per se that produce the looping effect. As the model is not a part of microeconomics, I will not consider it further.

[10]Hacking [1995, p.354] does talk about how characteristics such as being busy or selfish can be abstracted from people we know and summarised to form "profiles" or "personal inventories" that then become human kinds. However, the aim of microeconomics is not to describe the behaviour of selfish people, but to describe the behaviour of *everyone* in economic situations. Thus, the human kind created from the profile of selfishness does not capture the aim of the rationality assumption. One may also argue that homo economicus can be seen as a human kind. While I do not dispute this, I want to stress that homo economicus, while being derived from the rationality assumption, need not be implied by the assumption nor by any part of microeconomic theory.

effects were the overarching concept, it would therefore be impossible to describe the empirical effects we see of the rationality assumption by any of the theories discussed so far. If we do not want to exclude this phenomenon from the beginning, looping effects cannot be the overarching concept.

Third, Merton [1948] spends most of his paper considering how self-fulfilling prophecies can and have fostered discrimination [Merton, 1948, pp.196-210]. Using Hacking's later concept of human kinds and looping effects, Merton's discussion is a clear case of a looping effect. Since Merton initiated the debate on self-fulfilling prophecies, and since he exemplifies it with a looping effect, it makes more sense to see looping effects as a sub-genre of self-fulfilling prophecies (understood broadly as with the early thinkers, where scientific descriptions need not be predictions). Thus, while looping effects can be used to describe a subset of the possible ways science might change the world by describing it, they do not give us a general account of this phenomenon.

2.2.3 Performativity

The final group of theories - considering the phenomenon where knowledge of scientific descriptions changes the world - uses the concept *performativity*. Performativity is derived from Austin's definition of performative utterances [Austin, 1975].[11] A performative utterance is a speech act where the action a sentence describes, is performed by uttering that sentence, such as "I name the ship Aurora".

Performativity, as it is used here, was introduced by Callon [1998] in *The laws of the market*. It is specifically concerned with economics and describes the idea that economics can perform:

> economics, in the broad sense of the term, performs, shapes and formats the economy, rather than observing how it functions [Callon, 1998, p.2].

[11] For a discussion of the link between performativity and Austin's performative utterances, see Mäki [2013]. For a reply, see Guala [2016a]. I will return to this discussion in the end of this subsection.

When discussing what economics *does*, Callon refers to Latour's work on how science functions in action. Here, Latour argues that science is embedded in a larger societal network and that its practices can only be understood in such context [Latour, 1987]. Callon also refers to his own work on how to understand the progress of science [Callon, 1995]. Though neither of the texts mentions performative effects (or performation as Callon initially termed the phenomenon), it is worth noting that performativity is connected to the interrelation between science and the world, specifically how the progress of science can only be understood in the context of its interaction with and attempts to change the world. Performativity then, is originally concerned with (partially) intentional spillovers between science and world and how science is *used* and *developed* in order to change the world.[12]

The most widespread definition of performativity is given by MacKenzie [2008] in *An Engine, Not a Camera - How Financial Models Shape Markets*. Here, MacKenzie argues that

> Financial economics [...] did more than analyze markets; it altered them. It was an "engine" in a sense not intended by Friedman: an active force transforming its environment, not a camera passively recording it [MacKenzie, 2008, p.12].[13]

In order to analyse whether financial theory transformed finance, MacKenzie [2008, p.17] distinguishes between four types of performativity:

- **Generic performativity:** An aspect of economics (a theory, model, concept, procedure, data set, etc.) is used by participants in economic processes, regulators, etc.
- **Effective performativity:** The practical use of an aspect of economics has an effect on economic processes.
- **Barnesian performativity:** Practical use of an aspect of economics makes economic processes more like their depiction by economics.

[12] For an early discussion of performativity, see also Cochoy [1998], published in the same journal issue as Callon [1998] which was also made into the book *The laws of the Markets*.

[13] I will return to a discussion of Friedman [1953] in chapter 4.

- **Counterperformativity:** Practical use of an aspect of economics makes economic processes less like their depiction by economics.

Callon's and MacKenzie's definitions of performation and performativity are summarised in table 2.3. Note that for all definitions, the performative effects are conditioned on an aspect of economic theory *intentionally being used in practice*. Thus, performativity only occurs if an aspect of academic economics - like an algorithm, equation, or dataset - is being used by practitioners in the real world economy such as business employees, companies, policy makers, or regulators [MacKenzie, 2008, p.18]. Indeed, MacKenzie's book is concerned with the intentional application of financial theory to the market and whether this application changed the market so that it conformed more or less with the theory in question. Thus, the main difference between MacKenzie's four concepts of performativity is what effect the use of a theoretical element has on the real economy.

Interestingly, MacKenzie notes that Barnesian performativity could be seen as corresponding to Merton's self-fulfilling prophesies [MacKenzie, 2008, p.19]. However, he gives three reasons for not wanting to use this name for his concept. First, he stresses that Barnesian performativity and counter-performativity are subsets of the more general phenomenon (exemplified by the first two types of performativity) of incorporating economic theories into the infrastructure of markets. Second, MacKenzie argues that performativity is not concerned only with beliefs about or descriptions of a situation:

> While beliefs about markets are clearly important, an aspect of economics that is incorporated only into beliefs "in the heads" of economic actors may have a precarious status. A form of incorporation that is in some senses deeper is incorporation into algorithms, procedures, routines, and material devices. An economic model that is incorporated into these can have effects even if those who use them are skeptical of the model's virtues, unaware of its details, or even ignorant of its very existence [MacKenzie, 2008, p.19].

Third, MacKenzie notes that Merton's concept (and later interpretations of it) implies that beliefs can be made true by their dissemination alone. Contrary

Theorist	Concept	Initial situation	Change	Outcome	Domain	Main problem
Callon [1998]	Performation or performing.	An issue occurs in the real world.	Elements from economic theory are used to inform the solution.	By creating a solution based on the theory, the world comes to look more like the situation described by the theory.	Economics.	The role of science in the world. The effect of theories on society.
MacKenzie [2008]	Generic performativity.	An element from theory is used in the world.	No requirement of change in the world.	No requirement of an effect.	Economics.	The role of science in the world. The effect of theories on society.
MacKenzie [2008]	Effective performativity.	An element from theory is used in the world.	This changes some process in the world.	No requirement of an effect related to the theory.	Economics.	The role of science in the world. The effect of theories on society.
MacKenzie [2008]	Barensian performativity.	An element from theory is used in the world.	This changes the processes in the world.	The world looks more like the world described in the theory.	Economics.	The role of science in the world. The effect of theories on society.
MacKenzie [2008]	Counter-performativity.	An element from theory is used in the world.	This changes the processes in the world.	The world looks less like the world described in the theory.	Economics.	The role of science in the world. The effect of theories on society.

Table 2.3: **Summary of Callon's and MacKenzie's concepts of performativity.** The table summarises Callon's concept of performation along with Mackenzie's four concepts of performativity.

to this, MacKenzie argues that not all equations or theories could be made true in the market by virtue of sufficiently many (authoritative) people believing them [MacKenzie, 2008, p.20]. In order for an economic model to be applied to and influence the market, the model would - for example - have to have some favourable outcomes for the people who use it. Thus, when considering whether an economic theory is performative, we are not only looking at beliefs it might create in economic actors, but at technical transformations of institutions in the market because of the *application* of one or more theoretical elements. This focus on the technical transformation of external institutions due to the application of theoretical elements is something that has not been fully appreciated or considered in the earlier discussions of self-fulfilling prophecies.[14]

Despite MacKenzie's arguments that Barnesian performativity and self-fulfilling prophecies are not alike, the philosophical debate sparked by MacKenzie's book does seem to conflate MacKenzie's and Merton's definitions.[15] This is a problem, since performativity is concerned with the effects of a theoretical element being *used* in the real world whereas self-fulfilling prophecies are concerned with the effects of *disseminating* a belief, theory, or theoretical element. Thus, the two concepts do not necessarily describe the same phenomenon since some theoretical elements may not change the world by being disseminated (like the formulation of an algorithm), but may change the world if they are later *used* in the world (for example by being incorporated as a part of a computer program predicting asset prices). It may also be possible that some theoretical elements can change the world by their mere dissemination even though they are never actually *used* in the real world. An example of the latter could be the theory that people have no free will. Though it has been disseminated as a scientific (and philosophical) theory several times, it has - to my knowledge - not been used in practice for

[14]For a recent account of institutions - consistent with Bicchieri's account of social norms - see Guala [2016b].

[15]For papers equating performativity and self-fulfilling prophecies, see for example Felin and Foss [2009b,a], Ferraro et al. [2009], and Brisset [2016]. For papers focusing on the engineering claim of Callon's and MacKenzie's concepts see for example Santos and Rodrigues [2009], Santos [2011], and Curran [2018].

determining regulations or other guidelines concerning human conduct.

The conflation between self-fulfilling prophecies and performativity may have occurred because of the relative imprecise definitions of both terms [Guala, 2016a]. However, another possible source for the conflation is Callon's and MacKenzie's use of the word "performative". As already mentioned, performativity originates with Austin [1975] and is concerned with constitutive speech acts that change the world merely by being uttered (by the right people under the right circumstances). This focus on changes constituted by the mere utterance of a sentence (or theoretical element) sounds very similar to the idea considered for self-fulfilling prophecies that information about a theoretical element can cause a change in the world by changing people's beliefs and thereby their actions. Furthermore, beliefs are often disseminated by the exchange of information via the utterance or writing of a sentence. This again, suggests a connection between Austin's performative sentences and self-fulfilling prophecies. However, as I have already stressed, Austin's use of performative speech acts differs from Callon's and MacKenzie's account of performation and performativity, since the latter are concerned with the *use* of scientific elements and not the mere *utterance* of theoretical elements.

Finally, the use of "performativity" and its relation to Austin's *constitutive* speech acts has resulted in some confusion. Even though Callon and MacKenzie use performativity to consider *causal effects* between the use of a theoretical element and the economy, the association with Austin has sparked a debate in philosophy of science considering performativity's connection to anti-realism and the possible threat it poses to scientific realism [Lowe, 2018, Felin and Foss, 2009a,b]. This is unfortunate, since performativity does not pose a threat to realism when it is defined - as MacKenzie does - as a causal process between economics and parts of the economic reality:

> It is no threat to scientific realism about economics to acknowledge the possibility of causal economics-dependence of some items in the real-world economy. After all, economics as an academic discipline is itself social activity exercised within society, so such connections are a natural feature of social reality. Good social

> science will investigate such connections together with other causal connections in society at large [Mäki, 2012, p.21].

Thus, while performativity may be the most well-known concept at the moment, it suffers from being equated with self-fulfilling prophecies and from an unfocused debate, where many different problems - such as scientific realism, the extent of the phenomenon, how to test theories, and social implications - have been discussed at once.

In the next section, I discuss the different aspects of the above concepts - their similarities and differences - and present a new way of thinking about whether scientific descriptions can come to change the world. Finally, I discuss how changes caused by knowledge of the rationality assumption fit into the debate.

2.3 Thinking about changes

Looking at the three groups of theories above, we see that many different situations are discussed with regard to scientific descriptions that may change the world. We can get an understanding of the diversity of situations by considering the information in tables 2.1, 2.2, and 2.3. Looking at the columns labelled "initial situation", we see situations with i) beliefs, ii) information, iii) predictions, iv) categorisations, and v) practical use of elements of theories. Looking at the change that occurs, this is described as being due to i) dissemination of the prediction, ii) acting on beliefs or categorisations, or iii) implementing and changing technical practices and institutions in the world. Here, we also encounter a consideration of whether the effects are causal or constitutive. Finally, looking at the outcome, situations are described where there is i) a change in the truth value of beliefs or predictions (from true to false and vice versa), ii) a change in probability of the truth value, and iii) a change in similarity between world and prediction or categorisation. Thus, just by comparing the *situations* considered by the different concepts, we already see a wide diversity in the phenomena discussed.

In order to get a more precise understanding of how the situations discussed vary, I suggest to distinguish between them across seven dimensions:[16]

- First, "scientific descriptions" can be beliefs, predictions, categorisations, and theoretical elements.
- Second, changes can occur due to causal or constitutive effects.
- Third, changes can be related to dissemination of scientific descriptions, actions based on beliefs caused by the dissemination, and practical use of scientific descriptions.
- Fourth, the outcome can be related to changes in truth-value of scientific descriptions, changes in probabilities of descriptions being true or false, or changes in how similar the descriptions are to the world.
- Fifth, changes occurring in the world can be:

  a. Directly related to what the scientific theory is about.
  An example of such change is the change in the financial market to make it look more like the market described by the assumptions in the Black-Scholes equation [MacKenzie, 2008, ch.6].

  b. Concerning phenomena that are not the focus of the theory, but which are stated as auxiliary assumptions in the theory.
  An example of this is if the learning about the microeconomic theory of market equilibrium (clearing of markets due to supply and demand) can make people act more in accordance with the rationality assumption.

  c. Not related to the subject matter of the theory or to any auxiliary assumptions in the theory.
  One example of this is if the way the matching game is presented (men take the initiative to ask their most preferred woman on date, women can say "yes", "no", and "maybe") can affect how men

[16] The dimensions are derived on the basis of the situations and examples considered by the above theories. They should be seen as a heuristic help for thinking about the situations rather than as a set of necessary or sufficient conditions.

and women perceive their roles in society as being "active" and "passive" respectively.[17]

Another example is the claim some neuroscientists have made that there is no free will [Libet et al., 1993, Soon et al., 2008, Bode et al., 2011, Fried et al., 2011]. Learning this may affect people's proneness to cheat or help others, as several psychological experiments suggest [Vohs and Schooler, 2008, Baumeister and Brewer, 2012, Protzko et al., 2016].

- Sixth, the domains of science where the situations are thought to be able to occur can be all sciences, social sciences, scientific descriptions related to human affairs or groups, and economics.

- Seventh, the scope of the change can vary in degree from a change by one person or in one situation to a general change in society or by many people.

These dimensions can be combined in different ways, thus describing different situations. For example, self-fulfilling prophecies focusing on changes in truth value for predictions (fourth dimension) are typically combined with a talk about beliefs and predictions in the first dimension and to changes directly related to theory in the fifth dimension. However, the accounts vary on how they perceive the scientific domains that can cause such changes (dimension six) and they generally consider cases that have many different degrees of scope (dimension seven).

[17]The deferred acceptance algorithm, first proposed by Shapely and Gale, is one example of such a matching game [Gale and Shapley, 1962]. The algorithm is explained by Varian [2014, p.328] as follows: "The most famous algorithm, known as the deferred acceptance algorithm, goes like this. Step 1: Each man proposes to his most preferred woman. Step 2: Each woman records the list of proposals she receives on her dance card. Step 3: After all men have proposed to their most-preferred choice, each woman (gently) rejects all of the suitors except for her most preferred. Step 4: The rejected suitors propose to the next woman on their lists. Step 5: Continue to step 2 or terminate the algorithm when every woman has received an offer. This algorithm always produce a stable matching." Notice that the algorithm makes men the active agents whereas women are simply supposed to say "yes" or "no" to their potential proposers. This feeds into the gender roles in western industrialised societies where men are seen as active, aggressive, and spontaneous whereas women are seen as passive, weak, and responsive [Duncan, 1990, p.25].

The dimensions give us a structured way to talk about how the concepts and definitions of each concept differ. They also highlight the fact that there is a difference between the concepts and even internally to the definitions of each concept. In order to understand this difference, I suggest that we change focus from *how is each concept defined* to *what issue is the theorist concerned with*. By looking at the issues discussed, we see that the reason theorists focus on different dimensions is that they are concerned with different issues. We also see how the concepts have been confused because they have been used to consider the same issues, and how the same concept has been interpreted in many different ways because it has been used to consider many different issues. The last columns of tables 2.1, 2.2, and 2.3 give a short summary of the main issue each concept has been used to discuss. For the purpose of simplicity, I have divided the different issues into three headings, namely *methodological concerns in science*, *the function of science in society*, and *social implications of science*. In the next subsections I will clarify the debates related to the three headings. For each issue, I will specify how it relates to the current concepts in the literature as well as the proposed dimensions.

2.3.1 Methodological concerns in science

The idea that dissemination or use of scientific descriptions can come to change the world is relevant from a methodological perspective. If there exist cases where dissemination of a theory will change the world such that the truth value or truth tendency of the theory will change, then reliably testing *and* disseminating the theory will be difficult. This methodological concern is the main focus of Buck [1963], Romanos [1973], and Kopec [2011] when discussing reflexive predictions. It is also the main concern for Popper [1957] and Nagel [1961] who argue against the worry that social sciences cannot be objective or make predictions in the same way as natural sciences. The methodological concern of testing theories is also one of the issues debated by Hacking [1995]. Finally, we see a methodological concern in the performativity literature with regard to the question of realism versus anti-realism [Bergenholtz and Busch, 2016]. Thus, the methodological

consequences of the phenomenon have been raised and discussed for all three concepts.

When we consider methodological problems in science - for example the possibility of testing theories - some of the dimensions discussed above become salient. First, it is important whether changes occur due to the dissemination of information or due to the application of this information. This is important since it changes the scope of the problem (more scientific information is disseminated than applied). Second, it is relevant whether the truth values or truth tendencies of the predictions are changed, since this poses a direct problem for the ability to test theories or make predictions. Third, and for the same reason, it is relevant to consider whether the changes are directly related to the predictions of the theory. Thus, dimensions three, four, and five, are of special relevance when considering methodological issues. Further, it is relevant to discuss which scientific disciplines are affected by the phenomenon (dimension six), since this can say something about the severity of the methodological concern. The salience of these dimensions is in alignment with the debate found in the literature which tends to have focused on the domain of the problem, how easily it occurs, and how to overcome it.

By explicating that the phenomenon of changing the world via science dissemination is discussed *because* it poses a methodological problem for some or all sciences, it is possible to keep a clear structure of the debate and avoid entanglement with other issues considering the same phenomenon. Having an issue-oriented approach to the debate will also make it possible for theorists to see where the debates using the three different concepts overlap with regard to methodological concerns. This, in turn, can help theorists determining which papers to consider. Since it may be useful to use a specific term for the phenomenon when considering methodological questions, I suggest that *reflexive predictions* is used in this context. A reason for using this term is that the papers using *reflexive predictions* all explicitly have been concerned with whether the phenomenon poses a methodological problem in science. Next, I turn to a different issue, which this phenomenon has been used to consider, namely the interaction and interrelation between science

and society.

2.3.2 The function of science in society

The idea that the dissemination or use of scientific descriptions and elements can change the world, is also relevant when considering the interaction and interrelation between science and the world. It is obvious that *how the world is* matters for the scientific descriptions employed to talk about it. However, it is not always obvious that *scientific descriptions of the world* matter for how we change the world.

In order to consider the latter, we can first look at two examples where there is an obvious *application* of science - for example by underwriting policy interventions - which creates a change in the world. First, it is by now a well-established fact that smoking can cause lung cancer [Inoue-Choi et al., 2018, O'Keeffe et al., 2018]. Because of this, it is recommended that governments introduce a tax on cigarettes in order to reduce consumption [Ho et al., 2018]. In this way, scientific knowledge about the risks of smoking has contributed to the introduction of an additional tax on cigarettes and so to an increase in cigarette prices in many countries. Second, engineering provides us with a good example of how scientific knowledge is applied in the world. Here, knowledge of forces and materials enables us to construct bridges and skyscrapers. Again, it is not surprising that we use the knowledge obtained through science to improve our living conditions, institutions, and processes in our everyday life.

However, scientific elements may also be applied and used to change the world in less obvious ways. The case of stagflation occurring in the 1970s is an example of how the application of scientific knowledge can change the world in unforeseen ways. Stagflation is a situation where there is both high rates of inflation and a slow economic growth. From late 1930s until the 1970s Keynesian economics was the primary inspiration for economic policy. As a part of this theory, William Phillips published an empirical study on the connection between unemployment (correlated to growth of the economy) and the total money wage costs per production employee (which can be

related to inflation) [Phillips, 1962]. Though Phillips never made the claim, the empirical observations were taken to show a stable inverse relationship between inflation and unemployment. Since the relationship was assumed to be fixed, the curve was used to guide monetary policies until the 1970s where stagnation occurred. Thus, interventions based on the assumption in Keynesian economics that high inflation rates and slow economic growth cannot occur simultaneously changed the world to be less like its description in Keynesian economics [Lerner, 1977, Jahan et al., 2014].

Another example, of scientific knowledge having a (potentially) unforeseen effect on the world, is the example discussed by MacKenzie [2008] concerning options and option pricing in financial markets.[18] Before the 1960s, there was no market for financial derivative trading [MacKenzie, 2008, p.147]. In order to create such a market, legitimacy was needed. One way this was sought by the Mercantile Exchange was to ask Friedman to write a paper supporting the need of a future market in currency [Friedman, 1971].[19] This, Friedman agreed to, in exchange for $5000 [MacKenzie, 2008, p.147]. Here then, we already see an interplay between theoretical economics (or finance) and the real world.

Having established a market for financial derivatives - including options - traders needed to determine which options to buy and sell. The Black-Scholes-Merton equation for option pricing was one of several competing theories for how to determine whether an option is over- or undervalued. However, the equation started having an impact on the financial market immediately

[18] Options can be seen as small gambles on the change in value of a specific asset or stock (e.g. price of corn or the value of a dollar). A *call option* allows the owner to buy the underlying asset at a specific price within a certain time frame. To give an example: if you believe that the price of corn will rise within the next three months, you can buy a call option saying that you can by corn at the current price for the next three months. If the price rises more than what you paid for the option, you will have gained money. Similarly, a *put option* allows the owner to sell the underlying asset at a specific price within a certain time frame. Thus, if you believe the price of corn will fall, you can earn money by buying a put option, if the price of the option is less than the fall in corn price.

[19] A future is a contract where the owner is obliged to sell or buy a given asset at a specific price in a future time. Thus, it is a financial derivative like an option, but it differs since futures obliged the owner to sell or buy at the agreed time. In contrast, the option just gives the owner the opportunity, but does not require that the owner actually uses this opportunity.

[MacKenzie, 2008, pp.157-163]. First, the theoretical work surrounding the equation helped undermine the association between gambling and options. This was important for at least two reasons. A) There was a widespread hostility towards gambling, which was illegal in several U.S states [MacKenzie, 2008, p.15]. B) Since the crash in 1929, there had been a hostility towards options, since they can be interpreted as wagers on the price movement of stocks [MacKenzie, 2008, p.144]. Second, the equation was used to produce sheets of paper each month with the calculated prices of the different options. These sheets were then used on the trading floor by several traders [MacKenzie, 2008, pp.160-162]. Third, the Black-Scholes equation won popularity among traders because it was simple to understand the parameters and only one of the parameters (volatility) was unobservable [MacKenzie, 2008, pp.162-163]. Finally, the details of the equation were publicly available - something other equations for option pricing had not been [MacKenzie, 2008, p.163]. Because it was publicly available, it was possible for a large number of traders to adopt it and use it. According to MacKenzie [2008, p.166], the use of the equation caused a change in prices so that the prices predicted by the equation and the actual prices became increasingly closer between the 1970s and 1980s. Thus, the use of the Black-Scholes equation made the option prices more similar to the prices predicted by the equation.

Cases like this, where the application and use of science may not be obvious or may have unintended consequences, are relevant for understanding the interplay between sciences and the world. It is this focus that I suggest is the primary concern of Callon [1998] and MacKenzie [2008] when talking about performativity.

Considering the interplay between science and the world opens up a wide range of questions regarding how scientific ideas are implemented. These questions include (i) what the aim of the implementation is, (ii) what the scientific idea should look like in order to be implemented, (iii) how the idea may be implemented, (iv) how the implementation can obtain acceptance by the stakeholders, and (v) how to measure the success of the implementation. Using these questions, we see that the salient dimensions for the phenomenon

are the first (that we are concerned with scientific elements), the third (that it is the practical use of science that causes the change), and the fifth (how the changes relate to the theory that is implemented).[20]

We also see that it is important to specify the second dimension (whether the interplay is assumed to be causal or constitutive), since the latter opens up for the realism and anti-realism debate. When the second dimension is determined as causal, this tells us something about *how* and *why* the changes in the world occur: they occur because of an intentional *intervention* motivated by scientific elements.

Thus, as in the previous subsection, we see that the salient dimensions of the situation we consider change depending on the issue with which we are concerned. For example, the truth value of the scientific description (dimension four) is salient when we consider methodological challenges to science but it is not salient when we consider the function of science in society. Since the original definition and use of *performativity* are concerned with the interplay between science and society, I suggest that *performativity* is only used when considering this issue.

2.3.3 Social implications of science

Finally, the idea that scientific descriptions can change the world is relevant when we consider the social and normative implications of science. It is not surprising that science can have social implications - such as finding a cure for tuberculosis or sars-cov-2 - but some of the situations considered above point to the fact that the mere dissemination of scientific theories can change people's behaviour in unintended ways. An example of such unintended changes that can have social consequences is the changes related to the dissemination of microeconomic theory considered in this thesis.

As already discussed, both Merton [1948] and Hacking [1995] are concerned with the social implications of science. An example where a scientific categorisation had a social impact is the characterisation of the homosexual

[20] MacKenzie [2008, pp.21-25] is especially concerned with dimension five and how we can detect which influence the use of a financial theory has on the financial market.

man as more feminine than other men [Sandfort, 2005]. Same-sex intercourse has occurred throughout human history. However, it is only in the late 19th century that the concept of a homosexual person was defined and put under scientific scrutiny [Hacking, 1995, p.354]. In 1936, Terman and Miles published the first study considering the relation between sexuality and personality traits for American homosexual men [Terman, 1936]. The results, though generally drawing a more complex picture, were used to justify the myth that homosexual men are more feminine than other men [Sandfort, 2005]. The myth persisted in science at least until the late 1980s. In the same period, a masculinization of the male gay subculture occurred [Sandfort, 2005, pp.603,607]. Though there may be several independent reasons for why the masculinization occurred, it might have been related to the scientific conception of male homosexuals. If this is the case, then the scientific description of homosexual men as feminine is one reason that the group of men deliberately changed their expression so that the description became less correct.

The example of homosexuality is discussed by Hacking [1995]. However, Hacking's theory of human kinds and looping effects cannot account for all relevant phenomena when considering the social implications of sciences. This is because looping effects only account for descriptions and categorisations involving human kinds, that is, *groups* of people. Thus, it is relevant to distinguish between cases where a *group* is described and cases where *all* humans or agents are described. The latter is for example the case in the rationality assumption used in microeconomics, or when neuroscientists say that free will does not exist. Since *looping effect* is already used to consider changes in group behaviour, I suggest that this term is used when considering the implication of scientific descriptions aimed at groups. When considering the social implications of scientific descriptions aimed at *all* agents, I suggest that we use the term *self-fulfilling science* [Lowe, 2018]. I have four reasons for preferring this term. First, the term is associated with Merton who considers the social implication of science. Second, the word *self-fulfilling* indicates that the phenomenon is in some way about unintended changes caused by scientific descriptions. Third, the word *science* indicates that the phenomenon

is not limited to a certain scientific domain such as the social sciences. Finally, *self-fulfilling science* benefits from being a relatively new term, since it - to my knowledge - was first employed by Lowe [2018, p.344]. This means that it does not suffer from a history of different uses and definitions, wherefore it can more easily be applied to this specific issue without any confusion.[21]

Looking at the dimensions suggested earlier, we see that for debates considering the social implications of science, the first dimension (what does "scientific description" refer to) becomes salient. The scientific descriptions causing this phenomenon need not have an if-then form, as many of the theories considered in subsection 2.3.2 had. Instead, we see that the scientific descriptions under consideration can be described as predicative: all people or some people are X, act like X, or ought to act like X. Further, dimension three (is the change caused by dissemination of information, actions based on beliefs about the information, or use) is salient. This dimension is relevant when considering both the scope of the phenomenon and how to potentially avoid it. If we want to understand *how* and *why* these situations occur, a specification of how the changes relate to the theory (dimension five) may also be relevant. Finally, since the issue is related to a normative judgement about practices in science, dimension six (which domains of science the phenomenon can occur in) and seven (the scope of the change) are salient. Since the social implication of science is related to society, I suggest that the domain for considering this issue is sciences discussing the affairs of humans, intentional agents, and groups. Further, the scope of self-fulfilling sciences or looping effects is important in order to determine the extent of the social implications.

Considering the literature on the phenomenon that scientific descriptions can change the world, the issue of social implications seems to be the least debated topic (see tables 2.1, 2.2, and 2.3). Further there is a gap in the literature explaining the mechanisms through which scientific descriptions can cause unintended social implications for the agents they describe. Since scientific descriptions related to looping effects and self-fulfilling science can

[21] Lowe [2018] uses the effect of microeconomics on people's behaviour as the paradigmatic example of a self-fulfilling science.

be seen as predicative, the mechanisms through which they cause changes in society may differ from the mechanisms involved in direct intentional interventions. First, consider *looping effects*. Hacking [1995, pp.367-368,370] writes that looping effects occur due to the *moral connotation* of human kinds. Human kinds are kinds that people may *want* to be or *not want* to be [Hacking, 1995, p.367]. Though I will not defend it here, I suggest that scientific descriptions concerning groups of people may change the world by promoting a personal norm or by creating a group identity, potentially with social norms specific to that group.[22] Second, consider self-fulfilling science. The aim of this thesis is to argue that one way scientific descriptions can change the world is by promoting a social norm among the people they have been disseminated to. One example of a - potentially - self-fulfilling science is microeconomics and specifically its rationality assumption. I present my argument for how the rationality assumption can promote a social norm in chapter 6.

I end this chapter by considering how the claim that the rationality assumption can change the world relates to the seven dimensions, and how the argument that the rationality assumption can promote a social norm contributes to the literature discussed in this chapter.

2.3.4 Self-fulfilling rationality

Recall from chapter 1 that the different variants of the rationality assumption in microeconomics textbooks all assume everyone to have *rational preferences* and seek to *maximise* their *individual gains* (which can be defined as preferences, utility, expected utility, profit, or income). If the rationality assumption is a case of self-fulfilling science, it can be described as having the following characteristics on the seven dimensions proposed earlier:

- First, the rationality assumption provides propositional information about "human" behaviour in economic models.

[22] See Bicchieri [2006, pp.146-152] for one description of group identity.

- Second, dissemination of the rationality assumption is assumed to have a *causal* effect on the world.
- Third, the changes are related to dissemination of information about the rationality assumption and actions based on beliefs caused by this information.
- Fourth, the outcome can be related to changes in similarity between theory and world.
- Fifth, changes do not concern the phenomena that are the focus of the different models, but are instead related to the auxiliary assumptions of human behaviour used in the models.
- Sixth, the domain of the change is economic theory, especially microeconomics as it is taught at universities.
- Seventh, the scope of the change may increase due to the increased acceptance of neoclassical economics and teaching of it.[23]

By presenting an argument for how the dissemination and teaching of the rationality assumption can make it self-fulfilling, this thesis engages directly with the philosophical debate on self-fulfilling science. Specifically, it engages with the question of *how* dissemination and application of scientific theories can influence the world (part II). As I will discuss in chapter 3, some experimental papers have suggested that people's actions in laboratory experiments, along with the difference between economists and non-economists, may be explained by social norms [Gerlach, 2017, Baum et al., 2012, Knez and Camerer, 2000, Mulford et al., 2008, Peysakhovich and Rand, 2015]. With few exceptions, however, the current literature has not considered *how* these changes - for example due to social norms - occur. One of the few exceptions is Ferraro et al. [2005, p.9] who state that self-fulfilling science can occur if

a. theoretical elements are reflected in institutional design,

[23] I will return to this point in chapter 8.

b. the theories provide a language that can be used to comprehend the world, and

c. theories are accepted as true and as norms that govern behaviour.

They defend these explanations by describing how they apply to economics.

First, in order to show that theoretical elements can play a role in institutional design and thereby change the world, Ferraro et al. [2005, p.13] quote two examples. The first example is the use of the Black-Scholes equation discussed above. The second example shows how wage became the predominant goal for factory workers in the early twentieth century. It is not given that a high wage has to be a goal for workers [Schwartz et al., 1978, p.239]. However, because classical economics and Smith [1776] describe labour as a commodity, employers sought to utilize it as well as possible by completely specifying the tasks of each worker [Schwartz et al., 1978, pp.242-244]. This made it increasingly difficult for workers to negotiate other aspects of their work than wage. Thus, wage became the predominant goal of factory workers. Given the two examples, we see that the first way, Ferraro et al. [2005] consider how science can change the world, corresponds to cases of performativity, where changes occur through the *use* of theoretical elements.

Second, in order to support the claim that theories can change the world by providing a language through which to comprehend it, Ferraro et al. [2005, p.16] cite experiments from the priming and framing literature considering how changing the name of a game can change people's actions when they play the game. I will discuss these findings in section 3.5. Ferraro et al. conclude that normative language can have an effect on our behaviour, but they do not explain how.

Finally, Ferraro et al. [2005, p.14] argue that economic theory can be accepted as true by the people it is taught to and thereby create a norm for behaviour. They do this by referring to experimental literature - such as Marwell and Ames [1981] - indicating that economics education can influence behaviour. Unfortunately, their analysis of *how* teaching in economics creates a norm is inadequate for three reasons. First, they never define what they

mean by "norm". Second, they state that the rationality assumption is an assumption of self-interest that *predicts* how people act. This claim both misses the subtleties of the different variants of the rationality assumption and the point that the rationality assumption is an assumption about agent behaviour in economic models and not a prediction of how real people act. Third - though they hint that people might act self-interestedly to avoid looking foolish or naive - they do not give a full account of how teaching economics can cause people to believe that a norm of self-interest exists and should be followed. Thus, this thesis fills a gap in the current literature by providing a detailed argument for *how* microeconomics textbooks and teaching practices can change the world by promoting a social norm.

2.4 Conclusion

In this chapter I have discussed the larger philosophical project underlying the thesis. Looking at the current literature concerned with the phenomenon that scientific descriptions can change the world, we see that the debates surrounding the phenomenon can be divided into three main issues. First, philosophers have been concerned with the *methodological implications* for the social sciences. Second, philosophers and economic sociologists have used the phenomenon to discuss the *function of science in society*. Third, philosophers have considered the *social implications of science*. Using the proposed seven dimensions, we see that different dimensions of the phenomenon become salient depending on which issue one is interested in.

In order to avoid future confusion between the debates, I propose that we use the concept *self-fulfilling science* to discuss cases where scientific descriptions concerning all agents can have social implications by influencing behaviour. The aim of the thesis is to argue that it can occur, and to give a detailed account of one way it can do so. Thus, the thesis contributes to the literature by presenting one mechanism through which self-fulfilling science occurs.

To do this, I use the paradigmatic case of microeconomics and the dissemination of the rationality assumption through microeconomics textbooks.

Looking at the seven dimensions, we can get a coarse characteristic of the situation: the rationality assumption is presumed to have a causal effect on the world due to its dissemination and beliefs caused by it. As a result, human behaviour is allegedly changed to be more similar to the description of agent behaviour in microeconomics textbooks. The social implications of this phenomenon will depend on how many people are exposed to microeconomics textbooks and teaching as well as how one evaluates the consequences of the changed behaviour.

In the next chapter, I turn to the question of *whether* exposure to the rationality assumption through microeconomics textbooks and teaching practices can change people's behaviour. I do this by discussing the current empirical literature on the topic. As we shall see, the empirical findings support the hypothesis that exposure to microeconomic theory changes behaviour. In alignment with the argument of the thesis, they also suggest that the context of a situation may reduce or remove this difference.

Chapter 3

Empirical findings

3.1 Introduction

As already suggested by the studies of Marwell and Ames [1981], economics students act differently than non-economics students in some situations. In this chapter, I argue that the current empirical evidence supports the hypothesis that microeconomic theory is self-fulfilling due to the dissemination of the rationality assumption as it is described in microeconomics textbooks and university courses. Thus, the aim of the chapter is to argue *that* microeconomics is a case of self-fulfilling science. I do this in two steps. First, in sections 3.2, 3.3, and 3.4, I present the empirical literature studying the behaviour of economists (people who study or have studied economics) compared to non-economists (people who have not studied economics). These studies show that in situations discussed by microeconomics textbooks, economists act more in accordance with the textbook recommendations compared to non-economists. Second, in section 3.5, I argue that one way this behavioural difference can be caused is by framing a situation as economic, priming economic concepts, or - most importantly - disseminating microeconomics textbook excerpts describing the rationality assumption. Since the changes can be caused by disseminating economic theory, microeconomics qualifies

as a self-fulfilling science.

The literature relevant for the first part of my argument uses several different methods to test the difference in behaviour between economists and non-economists. These methods can be loosely divided into three categories. The first category is what I call *classroom experiments*. Classroom experiments refer to laboratory experiments that create an idealised and artificial environment in which it is possible to study controlled real-world behaviour [Morgan, 2012, p.279]. Classroom experiments, like the experiments conducted by Marwell and Ames [1981], can be used to study how people choose when they play one of the canonical games analysed in game theory. By making people play the games in a laboratory setting, it is possible to control how much information they are given when making their choices. Thus, classroom experiments can be used both to see how people choose when no information except payouts is given and whether their choices change if they are provided with more or varying information. The second category is what I call *survey experiments*. This category consists of experiments conducted via surveys to investigate how economists and non-economists choose when considering different dilemmas. The final category is what I call *observational studies*. Here, the real-world behaviour of economists and non-economists is observed. In the next three sections I analyse each of the categories in turn.

3.2 Classroom experiments

The games used to study behavioural differences between economists and non-economists in laboratory settings are typically *zero-sum games* or *social dilemmas* (a specific type of non-zero-sum games). A zero-sum game is characterised by a situation where one participant's loss or gain is exactly matched with other participants' gain or loss, respectively. A social dilemma is a type of game where there is a conflict between the interest of each individual player and the combined interest of all players. The prisoner's dilemma and the public goods game discussed in chapter 1 are examples of social dilemmas.

In this section, I first give a theoretical account of zero-sum games used

in the relevant classroom experiments as well as their game theoretical solutions. I then consider the experimental findings concerning the behaviour of economists and non-economists when playing zero-sum games. Next, I give a theoretical introduction to the social dilemmas used in these classroom experiments before turning to the empirical findings for each of the dilemmas. The findings support that economists act more in alignment with the rationality assumption compared to non-economists in classroom experiments. Further, they show that the behavioural difference may decrease when more information about the situation is provided or when participants are allowed to make promises to each other.

3.2.1 Zero-sum games - theoretical background

To my knowledge, three zero-sum games have been used to study economists' behaviour. These are the *dictator game*, the *ultimatum game*, and the *solidarity game*. Before describing the games and their game theoretical solutions, we need to introduce two additional solution concepts from game theory. Recall the discussion of a *strictly dominant strategy* from chapter 1. Informally, player i has a strictly dominant strategy if there exists one strategy that will always maximise player i's payoff regardless of what the other players choose to do. If player i is rational and has a strictly dominant strategy, player i will always choose that strategy. If all players in a game have a strictly dominant strategy, the solution to the game is given by the set of those strategies [Jehle and Reny, 2011, p.309].

However, games rarely have dominant strategies for all players. In these cases, we can employ a different solution concept, namely the *Nash equilibrium*. Recall from chapter 1 that a normal form game is specified by its players, each player's set of strategies, and the payoffs assigned to each player for each combination of strategies. We call a strategy, s_i, a *pure strategy*, if it is one of player i's strategies. In contrast, a *mixed strategy* for player i is defined as a probability distribution over i's set of pure strategies [Jehle and Reny, 2011, p.314]. Formally, we can specify a game, G, by: $G = (S_i, u_i)_{i=1}^{N}$ where S_i is the set of pure strategies for player i and u_i indicates the different

payoffs player i can receive depending on the choices of all players. Further, $S = S_1 \times S_2 \times ... \times S_N$ is the joint set of pure strategies for players $1, ..., N$. A pure strategy Nash equilibrium is defined as a set of strategies for all players, such that each strategy in that set is that player's best response to the other strategies in the set. Formally, Jehle and Reny [2011, p.312] defines a pure strategy Nash equilibrium as follows:

Definition 2 *Pure strategy Nash equilibrium*

Given a game, $G = (S_i, u_i)_{i=1}^N$, the joint strategy $\hat{s} \in S$ is a pure strategy Nash equilibrium of G if, for each player i, $u_i(\hat{s}) \geq u_i(s_i, \hat{s}_{-i})$ for all $s_i \in S_i$.

If there is only one pure strategy Nash equilibrium in a game, then that will be the game theoretical solution to the game, assuming that all players are rational.[1] Note that a pure strategy Nash equilibrium may not always exist. In such cases, players can consider using *mixed strategies*. Nash [1951] has shown that in a finite game - where each player has a finite number of moves - there is always at least one mixed strategy Nash equilibrium.

Finally, we can look at extensive form games, Γ, where the actions of the players are sequential. As before, a *strategy* of player i is defined as a total description of what player i will do in any possible situation that can arise in the game. Thus, a strategy for player i states what player i will do for all player i's decision nodes in the decision tree (assuming that the game can be represented diagrammatically). We can define a *subgame* of an extensive form game as the game following a decision node, x, where all players in the game know if x has been reached.

Given the definition of a subgame, we can now define a pure strategy subgame perfect Nash equilibrium as follows [Jehle and Reny, 2011, p.341]:

Definition 3 *Pure strategy subgame perfect Nash Equilibrium*

A joint pure strategy $s \in S$ is a pure strategy subgame perfect equilibrium of the extensive form game Γ, if s induces a Nash equilibrium in every subgame of Γ.

[1] Rationality is a necessary condition. For a list of sufficient conditions see section 5.3.

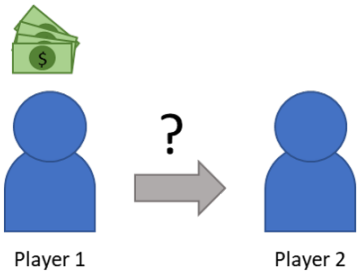


Figure 3.1: **Illustration of the dictator game.** Player 1 has to decide how to share the money they have received between themself and player 2.

These solution concepts enable us to examine the three zero-sum games in turn. For each game, I will describe its content and state the game theoretical solution to it. When describing the solutions, I assume that a rational player will always want to maximise their own monetary payoff. Though game theory need not make this assumption, it is typically assumed in experimental literature using games (see, for example, the studies considered in this chapter). Further, as we shall see in chapter 5, this assumption is often made and used in microeconomics textbooks.

Dictator game (DG): The game consists of two players. Player 1, *the dictator*, gets a set monetary amount and is told that they can divide it between themself and player 2 as they wish. Player 2, *the receiver*, receives whatever amount player 1 decides to give. The game is illustrated in figure 3.1.

Since player 2 cannot do anything, the game theoretical solution is only concerned with player 1. Assuming player 1 wishes to maximise their monetary payoff, player 1 will want to keep as much as possible for themself. Thus, the strictly dominant strategy for player 1 is to keep all the money for themself and give nothing to player 2.

Ultimatum game (UG): The game consists of two players. Player 1, *the proposer*, gets a set monetary amount and is asked to divide it between themself and player 2. Player 2, *the receiver*, is informed about the division proposed by player 1 and can either choose to accept or decline

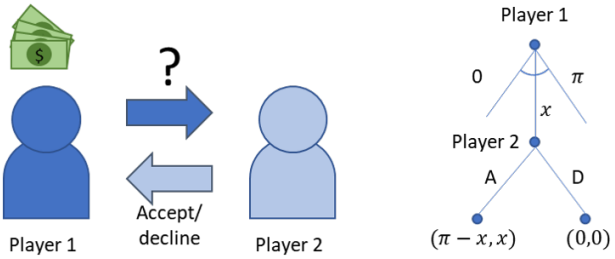


Figure 3.2: **Illustration of the ultimatum game.** *Left:* player 1 has to decide how to share the money they have received between themself and player 2. Player 2 can then accept or decline the offer. *Right:* a diagrammatic representation of the game.

the offer. If player 2 accepts, each player receives the proposed amount. If player 2 declines, neither player receives anything. The game is illustrated in figure 3.2.

In order to find the game theoretical solution to the ultimatum game, notice first that it consists of two subgames: the first subgame is the choice of player 1 and the second subgame is the choice of player 2. Let us start with the choice of player 2. Assume that player 1 gets the monetary amount π and has to decide the amount $x \in \pi$ to give to player 2. Now, player 2 has the option of accepting x and receiving x or rejecting x and receiving 0. This means that for all $x > 0$, player 2 will accept x, assuming that player 2 only cares about maximising their own monetary payoff in this one game. If $x = 0$, player 2 will be indifferent between accepting and rejecting the offer. Now consider the first subgame, where player 1 has to choose how to divide π. Player 1 knows that player 2 is rational and wants to maximise their own monetary payoff. Thus, player 1 knows that player 2 will accept any offer $x > 0$. Since player 1 also wants to maximise their monetary payoff, the best strategy for player 1 is to offer the smallest possible amount $\epsilon > 0$ to player 2. Thus, the unique subgame perfect Nash equilibrium to the ultimatum game is that player 1 offers player 2 the amount $\epsilon > 0$ and player 2 accepts, such that the outcome of the game will be a payoff of $\pi - \epsilon$ to player 1 and ϵ to player 2 [Rubinstein, 1982].

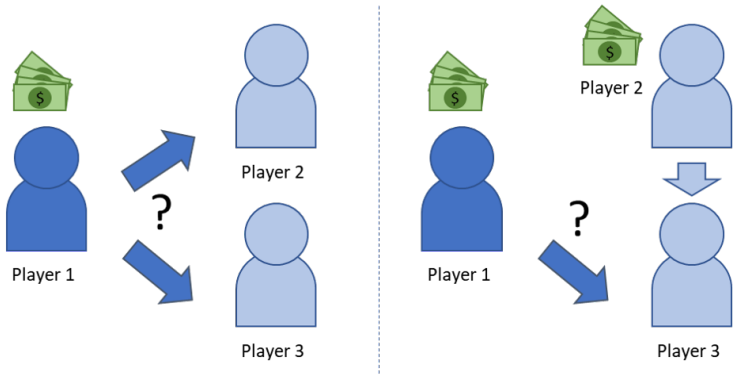


Figure 3.3: **Illustration of solidarity game.** Player 1 has to decide how to share the money - if they win - between themself and two losers in the group (*left*), or between themself and one loser in the group (*right*).

Solidarity game (SG): There are three players. Each player has an independent probability of $p = \frac{2}{3}$ for winning a set monetary amount. Before the players are informed of whether they have won, each player is asked - in the case they win - how much they want to give to the losing player(s) in the group i) when there are two losers and ii) when there is one loser. If everyone wins, no solidarity contributions will be made. The game is illustrated in figure 3.3.

The game theoretical solution to the solidarity game is the same as that for the dictator game, assuming that each player wishes to maximise their own monetary payoff. Since the two other players in the game are unable to respond to the offer made by a player, the strictly dominant strategy for each player is to keep everything for themself, in case they win the lottery.

3.2.2 Zero-sum games - empirical findings

Equipped with the knowledge from the previous subsection, we are now able to consider the experiments conducted to see if there is a behavioural difference between how economists and non-economists act in zero-sum games, and how this difference relates to the game theoretical solutions of the games. This subsection is divided into three. First, I review the findings

Study	Setting	Economic endorsement	Non-economists	Economists	Difference	Hypothesis Supported
Wang et al. [2011] Study 1	DG: the dictator can divide $10 as they wish.	Dictator keeps $10	Dictators keep $6.26 on average.	Dictators keep $7.80 on average.	$p < 0.001$***	**Yes**
Wang et al. [2011] Study 2	DG: the dictator can choose between a $9.25/$0.25 split or a $5/$5 split.	Dictator keeps $9.25.	40.1% of dictators keep $9.25.	64.1% of dictators keep $9.25.	$p < 0.05$*	**Yes**
Hole [2013]	DG: the dictator can divide the money in an investment pool as they wish.	Dictator offers 0% to the other player.	Dictators offers on average 46.3% to the other player.	Dictators offer on average 30.9% to the other player.	$p = 0.019$*	**Yes**
Gerlach [2017]	DG: the dictator can divide £12 as they wish.	Dictator offers £0.	Dictators offer £4.75 (arts) and 4.79 (science) on average.	Dictators offer £2.83 on average.	$p < 0.001$***	**Yes**
Ifcher and Zarghamee [2018]	DG: the dictator can divide $20 as they wish.	Dictator offers $0.	Dictators offer $5.59 on average.	Dictators offer $3.72 on average.	$p = 0.002$**	**Yes**

Table 3.4: **Summary of empirical findings for differences in behaviour when playing DG.** The table shows i) the study, ii) the experimental set-up, iii) the game theoretical prediction of behaviour in the game, iv) how non-economists behaved in the game, v) how economists behaved in the game, vi) whether the difference in the observed behaviour is significant, and vii) if the hypothesis that economists act more in accordance with the rationality assumption compared to non-economists is supported. * indicates a significance level of 0.05, ** indicates a significance of 0.01, and *** indicates a significance level of 0.001.

from the dictator game. Second, I review the findings for the ultimatum game, and, third, I review the findings from the solidarity game. Tables 3.4, 3.5, and 3.6 summarise the empirical findings.

Findings from dictator games

There are - to my knowledge - four studies that have considered how economic students choose in a dictator game compared to other students. The results of the four studies are summarised in table 3.4.

First, Wang et al. [2011] conducted a classroom experiment with two versions of the dictator game. In the first version, player 1 could divide $10 as they wished. In the second version player 1 could choose between dividing the amount equally (that is $5 per player) or keeping $9.25 and giving player 2 $0.25. The participants consisted of 67 economics students and 45 education students. In both versions, economics students kept significantly

more to themselves compared to education students ($p < 0.001$ and $p < 0.05$ respectively).

Second, Hole [2013] used a dictator game to consider the difference between economics students' and engineering students' conception of fairness. The results from the experiment showed that economics students made significantly smaller offers as dictators compared to engineering students, with $p = 0.019$ [Hole, 2013, pp.18-19]. Further, Hole [2013, pp.20-21] reports that the offers made by engineering students were largely in alignment with what they had previously stated to be fair: 77.6% offered the exact amount they had stated as fair and only 5.2% offered nothing to the other player. In contrast, 43.5% of economics students made the offer they had originally indicated as fair, and 26.6% of the economics students chose to offer nothing to the other player. The remaining participants offered something to the other player, but less than what they initially had indicated as a fair offer.

Third, Gerlach [2017] also used a dictator game to consider economics students conception of fairness. The participants in the study were either economics students, art majors, or science majors. Each game had three players and consisted of two stages. The first stage was a normal dictator game with player 1 as dictator and player 2 as receiver. Player 1 was given an amount of \$12 to divide between themself and player 2. In the second stage, a third-party judge (player 3) was able to either accept the proposed division or to decline the proposed division and instead make a new division of the \$12 between player 1 and player 2. If player 3 chose to accept the division, they would receive \$7. If they decided to decline the division, they would receive \$5. The results from the first stage of the game showed that economics students on average made significantly smaller offers as dictators compared to both arts majors and science majors with $p < 0.001$ in both cases [Gerlach, 2017, p.6]. The results from the second stage of the game further showed that economics students who had the role as judges (player 3) were significantly less likely to veto offers that they perceived as unfair compared to the other students ($p = 0.004$) [Gerlach, 2017, table 6, p.12].

Finally, Ifcher and Zarghamee [2018] conducted an extensive experiment where the first phase asked participants to play four different games from

Study	Setting	Economic endorsement	Non-economists	Economists	Difference	Hypothesis Supported
Carter and Irons [1991]	UG: the proposer can divide \$10 as they wish.	Proposer keeps \$9.5. Receiver accepts all above \$0.	Proposers keeps \$5.44 on average. Receivers minimum accept \$2.44 on average.	Proposers keep \$6.15 on average. Receivers minimum accept \$1.70 on average.	$p < 0.025$* $p < 0.025$*	**Yes**
Ifcher and Zarghamee [2018]	UG: the proposer can divide \$20 as they wish.	Proposer keeps \$19. Receiver accepts all above \$0.	Proposers keeps \$8.88 on average. Receivers minimum accept \$4.61 on average.	Proposers keep \$7.04 on average. Receivers minimum accept \$3.72 on average.	$p < 0.001$*** $p = 0.091$	**Yes**

Table 3.5: **Summary of empirical findings for differences in behaviour when playing UG.** * indicates a significance level of 0.05 and *** indicates a significance level of 0.001.

traditional game theory.[2] Looking at the results from the dictator game, Ifcher and Zarghamee [2018, p.60] report that participants with previous exposure to game theory gave significantly less to the other player compared to students with no previous exposure ($p = 0.002$).[3]

Findings from ultimatum games

Turning to the empirical results on behavioural differences for the ultimatum game, we have two sources. The first source is Carter and Irons [1991, p.174]. They used a normal UG to consider the differences between economics students and non-economics students. The proposer was given \$10 to divide between themself and the receiver, with a minimum division unit of \$0.5. As indicated in table 3.5, economics students on average kept \$6.15 as proposers and were on average willing to accept offers down to \$1.70. In contrast, non-economics students on average proposed to keep \$5.44 and were willing to accept offers down to \$2.44 on average. The difference in both proposal size and minimum offer acceptance are significant between the two groups, with economists acting more in alignment with the game theoretical solution.

Ifcher and Zarghamee [2018] also used an ultimatum game to test the differences between participants with previous exposure to game theory and

[2]For a full description of their experiment, see section 3.5.

[3]Looking at whether the participants had taken any prior economics courses, the difference in behaviour between economists and non-economists is significant with $p = 0.012$ with economists giving less than non-economists [Ifcher and Zarghamee, 2018, p.62, table 12].

Study	Setting	Economic endorsement	Non-economists	Economists	Difference	Hypothesis Supported
Selten and Ockenfels [1998]	SG: each player chooses how to divide 10DM if one or two losers.	The player keeps 10DM in both cases.	Players give 2.84DM if one loser and 1.78DM if two losers.	Players give 1.49DM if one loser and 0.89DM if two losers.	$p = 0.0014^{***}$ $p = 0.0012^{***}$	**Yes**

Table 3.6: **Summary of empirical findings for differences in behaviour when playing SG.** *** indicates a significance level of 0.001.

participants with no previous exposure to game theory. Here, the proposer was allowed to divide $20 as they wished, with a minimum division unit of $1. Looking at how much the proposers offered, there is a significant difference between the two groups, with participants in the former group offering less to the receiver ($p < 0.001$). However, the difference in minimum offer acceptance is not significant between the two groups [Ifcher and Zarghamee, 2018, p.60,62].[4]

Findings from the solidarity game

Finally, Selten and Ockenfels [1998, p.529] considered the behaviour of economics students compared to other students in a solidarity game. They report that there is a significant difference between economists and non-economists. Since they do not state what the difference is, I have reconstructed their data set based on Selten and Ockenfels [1998, table 6, appendix B, pp.534-537]. Using a one-tailed Mann-Whitney U-test we see that economists give less than non-economists in the case of two losers ($p = 0.0012$) and in the case of one loser ($p = 0.0014$). The results are reported in table 3.6.[5]

[4]When looking at exposure to prior training in economics, Ifcher and Zarghamee [2018, p.62, table 12] did not find a significant difference between proposers ($p = 0.912$) or receivers ($p = 0.716$).

[5]Selten and Ockenfels [1998] primarily focus on difference in behaviour between economists and non-economists based on gender. Here, they report a significant difference between male economists and male non-economists in the two cases ($p < 0.004$ and $p < 0.002$ respectively). They further say that there is no significant difference between female economists and female non-economists [Selten and Ockenfels, 1998, p.530]. (I have calculated the significance levels to be $p = 0.43$ and $p = 0.49$.) From these results, they argue that there *only* is an education effect for male economists, and not for female economists. Though there clearly is a general gender effect in this study [Selten and Ockenfels, 1998, p.529], it is worth noticing that the difference in significance between male and female economics students may be due to the very small sample size of female economists: out of

Summarising the experimental findings from zero-sum games, the results consistently show a significant difference in behaviour between economists and non-economists. The only exception is the result reported by Ifcher and Zarghamee [2018] in the ultimatum game, where there was no significant difference in minimum acceptance rate for receivers. However, since Ifcher and Zarghamee [2018] do find a significant difference between participants with and without previous exposure to game theory when they are proposers, the hypothesis that economists act more in accordance with the endorsements of microeconomic theory is still supported.

3.2.3 Social dilemmas - theoretical background

The other type of games considered in the experimental literature on economists' behaviour falls within a broad family of games known as social dilemmas. Social dilemmas are games characterised by a conflict of interests between the individual player and the collective. Specifically, if all players choose what is best for themselves, it will lead to an outcome where everyone is worse off compared to what they could have been. Using the terminology from game theory, a social dilemma is a game where a pure strategy Nash equilibrium in the game will result in a Pareto suboptimal outcome.

Pareto optimality was shortly discussed in chapter 1. It is a measure of efficiency and is therefore also referred to as *Pareto efficiency*. A more stringent definition of the term is given by Jehle and Reny [2011, p.183]:

Definition 4 *Pareto efficient allocation or outcome*

In general, when it is possible to make someone better off and no one worse off, we say that a *Pareto improvement* can be made. If there is no way at all to make a Pareto improvement, then we say that the situation is *Pareto efficient*.

Going back to the terminology of Pareto *optimality* (as is preferred within game theory), an outcome is *Pareto suboptimal* if it is not a Pareto optimal or

the 120 participants, 17 participants were male economists while only 6 participants were female economists [Selten and Ockenfels, 1998, appendix B, pp.534-537.]. This possible uncertainty is not discussed in the paper.

Pareto efficient outcome. The problem of social dilemmas is that the Pareto optimal outcome typically lacks the stability property of a Nash equilibrium. Thus, if the players in a social dilemma are rational, they will choose the equilibrium solution rather than the strategies that will result in a Pareto optimal outcome.

To my knowledge, four different social dilemmas have been used to study economists' behaviour. The games are *public goods game*, *prisoner's dilemma*, *stag hunt*, and *trust game*. Before introducing each of the four games and their game theoretical solutions, however, we need to define two additional solution concepts from game theory. The first solution concept is that of a *payoff dominant Nash equilibrium* [Harsanyi et al., 1988].

Definition 5 *Payoff dominant Nash equilibrium*

A Nash equilibrium is payoff dominant if it is Pareto superior to all other Nash equilibria in the game. If there is only one Nash equilibrium in the game, this equilibrium is trivially also a payoff dominant Nash equilibrium.

The second solution concept is that of a *risk dominant Nash equilibrium* [Harsanyi, 1995, p.92].

Definition 6 *Risk dominant Nash equilibrium*

A Nash equilibrium is risk dominant if it is the least risky Nash equilibrium in the game. If there is only one Nash equilibrium in a game, this is trivially a risk dominant Nash equilibrium.[6]

In cases where there is more than one Nash equilibrium, and one pure strategy Nash equilibrium is payoff dominant and another is risk dominant, the question arises of which Nash equilibrium rational players will play. Initially, Harsanyi et al. [1988] suggested that rational players would play the payoff dominant Nash equilibrium. However, persuaded by the arguments of Aumann [1990], Harsanyi [1995] later suggested that a rational player would play the risk dominant strategy in a competitive game (such as the

[6] Harsanyi [1995] uses a multilateral risk dominance measure to decide which equilibrium is the least risky. This is opposed to the bilateral risk measure used by Harsanyi et al. [1988]. For as technical description of the solution concept, see Harsanyi [1995].

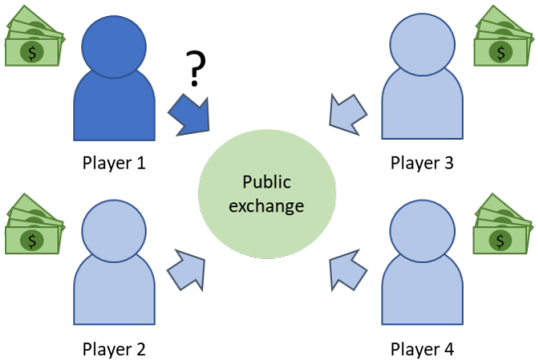


Figure 3.7: **Illustration of public goods game.** Each player has to decide whether to invest in an individual exchange or a public exchange. The exchange rate is highest for the public exchange, but everything invested in the public exchange will be shared equally among all players, regardless of who invested it.

ones we are considering here). Further, recent findings in evolutionary game theory - where a game is played repeatedly in a population of N players - suggest that if the initial strategy is randomly chosen among the players, the risk dominant strategy is most likely to become the fixed strategy (the strategy that most players play) [Sandholm, 2009, p.4].

Given the above solution concepts, we can now turn to a description of the four social dilemmas.

Public goods game (PGG): The number of players in a PGG can vary. However, a commonly used version of the game has four players. In the game, each player has an initial monetary endowment and is asked to divide the amount between an individual exchange and a public exchange. The money invested in the individual exchange is returned to the player with an exchange rate, x. The money invested in the public exchange is summed for the four players and increased with another exchange rate, y, where $y > x$. The money in the public exchange is then shared equally between the four players, regardless of who invested in the public exchange. Figure 3.7 illustrates the game.

As discussed in chapter 1, the game theoretical solution to a public goods game - assuming that all players want to maximise their monetary

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	(R,R)	(S,T)
	Defect	(T,S)	(P,P)

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	(2,2)	(0,3)
	Defect	(3,0)	(1,1)

Figure 3.8: **Matrix representation of prisoner's dilemma.** *Left:* the generic matrix representation of prisoner's dilemma with $T > R > P > S$. *Right:* an example of concrete payoffs satisfying the conditions for a prisoner's dilemma game.

payoffs - is that no player will invest in the public exchange. The public goods game is a social dilemma since the only Nash equilibrium in the game (no player invest in the public exchange) will result on a Pareto suboptimal outcome.

Prisoner's dilemma (PD): The game typically has two players with symmetric payoffs. Each player can either choose to cooperate or to defect. The players choose simultaneously and without knowledge of each other's choices. If both players choose to cooperate, they will each get a payoff R (the reward payoff). If one player chooses to cooperate and the other to defect, the first will receive S (sucker's payoff) and the other T (temptation payoff). Finally, if both players choose to defect, they will each receive P (punishment payoff). The payoff structure is such that $T > R > P > S$. If prisoner's dilemma is played as an iterated game (where the players play several rounds of prisoner's dilemma against the same people) the further requirement of $\frac{T+S}{2} < R$ is used. The game structure is depicted in figure 3.8.

Recall the discussion of PD from chapter 1: if player 2 chooses to cooperate, player 1 will earn the most by defecting, since $T > R$. Further, if player 2 chooses defect, player 1 will earn most by defecting, since $P > S$. Since the game is symmetrical, player 2 will likewise always earn the most by defecting. Thus, the dominant strategy for both players is to defect, resulting in the Nash equilibrium (defect,defect). However, if both players defect, they will each earn P, which is less than they would have earned, had both players cooperated ($R > P$).

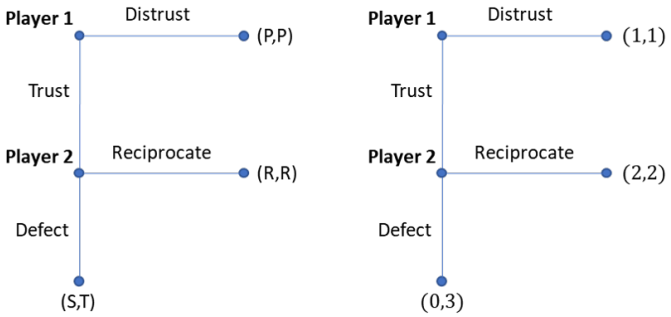


Figure 3.9: **Diagrammatic representation of a trust game.** *Left:* the generic diagrammatic representation of the trust game with payoffs $T > R > P > S$. *Right:* an example of concrete payoffs satisfying the conditions for a trust game.

Trust game (TG): The game typically has two players. The game has the same payoff structure as a prisoner's dilemma with $T > R > P > S$. However, instead of letting the two players choose simultaneously, the choices in a trust game are sequential. Thus, player 1 can first choose between distrusting and trusting. If player 1 chooses to distrust player 2, both players will get P. Alternatively, player 1 can choose to trust player 2. In this case, player 2 gets the option between reciprocating, so that both players receive R, or defecting, in which case player 2 receives T while player 1 receives S. The game is called the trust game, because player 1 has to decide whether to trust that player 2 will reciprocate their trust or take the larger amount for themself. The game is illustrated in figure 3.9.

The game theoretical solution to the trust game can be found as follows. First, consider the choice for player 2. Player 2 can either reciprocate and receive R or defect and receive T. Assuming that player 2 wants to maximise their monetary payoff, player 2 will choose to defect since $T > R$. Now consider the choices for player 1. Assume that player 1 wants to maximise their monetary payoff and knows that player 2 wants the same. Player 1 will then have the choice between distrusting and receiving P, or trusting and receiving S. Since $P > S$, player 1 will choose to distrust and end the game at once.

		Player 2	
		Stag	Hare
Player 1	Stag	(R,R)	(S,T)
	Hare	(T,S)	(P,P)

		Player 2	
		Stag	Hare
Player 1	Stag	(3,3)	(0,2)
	Hare	(2,0)	(1,1)

Figure 3.10: **Matrix representation of stag hunt.** *Left:* the generic matrix representation of stag hunt with $R > T > P > S$. *Right:* an example of concrete payoffs satisfying the conditions for a stag hunt game.

Stag hunt (SH): The story accompanying this game is a story about a hunt, first told by Rousseau [Skyrms, 2001, p.30]:

> If it was a matter of hunting a deer, everyone well realized that he must remain faithful to his post; but if a hare happened to pass within reach of one of them, we cannot doubt that he would have gone off in pursuit of it without scruple [...]

Simply put, the idea is that the hunters can either cooperate to capture the larger prey (the stag), or they can each decide to catch a smaller prey (a hare). In order to capture the stag, all hunters have to choose to do so. Thus, if one hunter decides to go for a hare, the other hunters aiming for the stag will be unable to catch it. However, each hunter is able to catch a hare on their own, without the help of the other hunters.

In game theory, the situation is typically translated into a game with two players. The players can either choose the stag or the hare. The players choose their strategies simultaneously and without knowing what the other player will choose. The payoff structure differs from the PD by having $R > T$ such that $R > T > P > S$. The game is illustrated in figure 3.10.

The change in payoff structure, making $R > T$, means that there are two pure strategy Nash equilibria in a stag hunt game. To see this, consider player 1's options. If player 2 chooses hare, player 1's best strategy is also to choose hare, since this will give player 1 a payoff of P rather than S and $P > S$. Since the payoffs are symmetric, player 2's best response to player 1 choosing hare is also to choose hare, and thus

we have a Nash equilibrium (hare,hare). However, if player 2 chooses stag, player 1's best response will be to also choose stag, since that will give them a payoff of R rather than T and $R > T$. Since the payoffs are symmetric, the same is the case for player 2, and thus we have a second pure strategy Nash equilibrium (stag,stag).[7]

The fact that there are two pure strategy Nash equilibria in stag hunt means that the game is a coordination game, where the players will be better off by choosing the same strategies. In order to determine the game theoretical solution to the game, we thus have to determine which of the two pure strategy Nash equilibria rational players will choose. Looking first at the equilibrium (stag,stag), we see that this Nash equilibrium is payoff dominant, since $R > P$. Looking at the Nash equilibrium (hare,hare), however, we see that this equilibrium is risk dominant when $T + P \geq R + S$, which, for example, is the case in the example given to the right in figure 3.10.[8] Recalling the discussion at the beginning of this section, there are good reasons to assume that rational players seeking to maximise their monetary payoffs will choose the risk dominant Nash equilibrium (hare,hare). Thus, stag hunt is also a social dilemma since it has a Nash equilibrium that, when chosen, will result in a Pareto suboptimal outcome.

3.2.4 Social dilemmas - empirical findings

Equipped with the above theoretical knowledge of social dilemmas, we are now ready to turn to the experimental literature considering how economists and non-economists choose in these situations. The subsection is divided in accordance with the different games studied in the literature. Thus, I will first report the findings from experiments using public goods games. Second, I will turn to experiments conducted with prisoner's dilemma. Finally, I will

[7] The game further has a mixed strategy solution, where the exact solution will depend on the payoffs. For simplicity, I will here restrict my analysis to the two pure strategy Nash equilibria.

[8] The condition is arrived at by using Harsanyi et al. [1988, lemma 5.4.4] rather than the measure laid out in Harsanyi [1995].

Study	Setting	Economic endorsement	Non-economists	Economists	Difference	Hypothesis Supported
Marwell and Ames [1981]	PGG: each player can invest 225 tokens as they wish.	No tokens are invested in the public exchange.	Players invest 42% on average in the public exchange.	Players invest 20% on average in the public exchange.	$p < 0.05$*	**Yes**
Ifcher and Zarghamee [2018]	PPG: each player can invest \$20 as they wish.	No player invests in the public exchange.	Players invest \$7.57 on average in the public exchange.	Players invest \$5.80 on average in the public exchange.	$p = 0.017$*	**Yes**

Table 3.11: **Summary of empirical findings for differences in behaviour when playing PGG.** * indicates a significance level of 0.05.

consider the experiments using stag hunt and trust games. For each part, I will present a table summarising the experimental findings and stating whether they are in alignment with the hypothesis that economists act more in alignment with the game theoretical solutions compared to non-economists.

Findings from public goods games

To my knowledge, there are two studies that have used a PGG to consider the difference in behaviour between economists and non-economists. Their results are summarised in table 3.11.

The first study is the one conducted by Marwell and Ames [1981], which has already been discussed in chapter 1. The study consisted of 12 experiments, of which the 12th was done with economics students. Participants were usually given 225 tokens to invest in the public or individual exchange. Marwell and Ames [1981] found a general willingness among non-economists to contribute between 40% and 60% of their tokens to the public exchange. For economics students, on the other hand, the average investment in the public exchange was 20%. The difference is significant with $p < 0.05$ compared to the control experiment which had an average contribution of 42% to the public exchange [Marwell and Ames, 1981, p.307]. A summary of all 12 experiments can be seen in table 1.2, chapter 1.[9]

[9]It should be noted that experiment 12, using economics students as participants, also increased the exchange rates of the individual and the public exchange with a factor of two. As seen in experiment six of the study, increasing the exchange rates with a factor five can also decrease the investments in the public exchange (see table 1.2). However, the decrease observed in experiment six is much smaller than the decrease observed in experiment 12.

The second study is conducted by Ifcher and Zarghamee [2018] who also used a PGG in their experiment. Here, participants were given $20 to invest in an individual or public exchange. Ifcher and Zarghamee [2018, pp.60,62 table 12] found that participants with prior training in economics on average invested less in the public exchange compared to participants with no prior economics training ($p = 0.017$).[10]

Thus, when considering the behaviour of economists and non-economists in a PGG, the results support the hypothesis that economists behave more in accordance with the solutions endorsed by game theory compared to non-economists.

Findings from prisoner's dilemma

Turning to the prisoner's dilemma, I have found four studies considering the behavioural differences between economists and non-economists when playing the game. The results from the four studies are summarised in table 3.12.[11]

The latest study, conducted by Ifcher and Zarghamee [2018, p.62, table 12] reports a significant difference ($p = 0.004$) in the amount of defection in a PD with students who have prior training in economics defecting more than students who do not have prior training in economics. The result is also significant ($p = 0.026$) when analysing participants with and without previous exposure to game theory [Ifcher and Zarghamee, 2018, p.62, table 11].

Frank et al. [1993] conducted another study consisting of three experiments, where different participants played PD under varying circumstances. In

Thus, this additional change cannot explain away the difference in behaviour between economists and non-economists.

[10] Ifcher and Zarghamee [2018, pp.60,62 table 11] also found that if participants were divided in accordance with previous game theoretical exposure, the average difference in investment is not significant ($p = 0.270$). 72 participants out of 276 said that they had studied game theory while 100 participants out of 276 said that they had taken economics courses, thus one reason for the difference in significance may be that the former is a subset of the latter, and that the 28 participants, who have studied economics but not game theory, already act as predicted by economic theory, and thus have biased the results.

[11] See chapter 7, subsection 7.2.2, and Buchter et al. [2020] for a fifth experiment that also shows a behavioural difference between economists and non-economists.

Study	Setting	Economic endorsement	Non-economists	Economists	Difference	Hypothesis Supported
Ifcher and Zarghamee [2018]	PD: players can choose to defect or cooperate.	Both players defect.	53% of players defect.	71% of players defect.	$p = 0.004$**	**Yes**
Frank et al. [1993]	PD: 10 or 30 minutes prior interaction.	Both players defect.	38.8% of players defect.	60.4% of players defect.	$p < 0.005$**	**Yes**
Hu and Liu [2003]	PG: prior interaction with promises.	Both players defect.	49.62% of players defect.	34.84% of players defect.	$p < 0.001$***	**No**
Ahmed [2008]	PD: players can choose to defect or cooperate.	Both players defect.	65% of players cooperate.	38.3% of players cooperate.	$p = 0.003$**	**Yes**

Table 3.12: **Summary of empirical findings for differences in behaviour when playing PD.** * indicates a significance level of 0.05, ** indicates a significance level of 0.01, and *** indicates a significance level of 0.001.

the first experiment, participants were divided into groups of three and were able to talk for 10 minutes prior to playing the game against their group members. In the second experiment, the time spend in the group was increased to 30 minutes. Finally, in the third experiment, participants were told that - in contrast to the two previous experiments - they were allowed to make promises to each other regarding their strategy choices. Participants were also informed that their responses to the PD would be anonymous and that the other group members would not be able to know what they actually did. Controlling for gender and age, Frank et al. [1993, p.164] found that economics majors defected significantly more than non-economics majors when the results from the three experiments are considered together ($p < 0.005$). Looking at the experiments separately, however, the difference disappears when participants are allowed to make promises [Frank et al., 1993, p.166].[12]

Hu and Liu [2003] also find that economists do not defect more than non-economists when participants are allowed to make promises prior to playing a PD. In their experiment, participants were also divided into groups of three and had the opportunity to talk and make promises for 20 minutes prior to playing a PD. Hu and Liu [2003, p.691] find that economics majors

[12]It is significant for the two other experiments at a $p < 0.01$ level.

cooperate more compared to non-economics majors ($p < 0.001$). The finding that economists tend to cooperate more than non-economists when promise-making is allowed is an unusual discovery that makes this study a rare outlier in the empirical literature on economists' behaviour compared to non-economists. The findings of Hu and Liu combined with the results from Frank et al. [1993] indicate that promise-making can remove the behavioural differences between economics students and non-economics students.

Finally, Ahmed [2008] conducted an experiment using a one-shot PD. Participants were either from humanities, economics, or police cadets. The participant pool was further divided into freshmen and seniors for each group. Ahmed [2008] only analyses the participant pool when it is split into the six subgroups. There are 30 participants per subgroup. Ahmed [2008, p.304] found no significant difference between freshmen economists and other freshmen (a χ^2-test for all three groups has $p = 0.379$). For *seniors*, however, there was a significant difference between economists where 30% cooperated, and police cadets where 87% cooperated. Ahmed [2008] explains this difference by appealing to an increase in cooperation for police cadets ($p < 0.08$ which he describes as significant despite it being above 5%) rather than the behaviour of economics students [Ahmed, 2008, p.304]. Based on this, Ahmed [2008] concludes that his results do not suggest that economists' behaviour differ from other people's behaviour in general.

Using the results reported by Ahmed [2008], I conducted an analysis considering the behaviour of freshmen *and* senior economics students, compared to the control group of freshmen *and* senior humanities students [Ahmed, 2008, p.304, table III]. Combining the freshmen and senior participants for each education does not bias the results since there are equally many freshmen and seniors in each group. Analysing the two groups, we see that 65% of the humanity students cooperate, while only 38.3% of the economics students cooperate. Using a χ^2-test, I find that the difference between the two groups is significant with $p = 0.003$. Thus, the results reported in the experiment *do* show a significant difference in behaviour between economists and non-economists, even when police cadets (who are reported to cooperate more than humanities students) are not considered. The result reported in

Study	Setting	Economic endorsement	Non-economists	Economists	Difference	Hypothesis Supported
Ahmed [2008]	SH: players can choose stag or hare.	Both players choose hare.	76.7% of players choose stag.	53.3% of players choose stag.	$p = 0.007$**	**Yes**
Haucap and Müller [2014]	TG: player 1 can choose to trust or distrust. Player 2 can choose to reciprocate or defect.	Player 1 distrusts. Player 2 defects.	53% of player 1s trust. 51% of player 2s reciprocate.	41% of player 1s trust. 40% of player 2s reciprocates.	$p = 0.004$**	**Yes**

Table 3.13: **Summary of empirical findings for differences in behaviour when playing stag hunt and the trust game.** ** indicates a significance level of 0.01.

table 3.12 is the result from the more general analysis that I conducted.

For prisoner's dilemma then, there is evidence that economists act more in accordance with the game theoretical solution compared to non-economists [Ifcher and Zarghamee, 2018, Frank et al., 1993, Ahmed, 2008]. Further, we see that the difference between the two groups disappears when participants are allowed to make promises before playing the game [Frank et al., 1993, Hu and Liu, 2003]. Finally, Ahmed [2008] suggests that it is also possible to make people cooperate more by giving them a team-focused training (as is the case for the police cadets compared to the humanities students in his experiment).

Findings from stag hunt and trust games

To my knowledge, there has been one classroom experiment testing behavioural differences between economists and non-economists using a stag hunt game and one classroom experiment using a trust game. Here, I consider the two in turn. Table 3.13 provides a summary of the experimental results.

Ahmed [2008] also asked participants to play stag hunt. Again, Ahmed [2008, p.304, table IV] argues that the difference between junior police cadets and senior police cadets ($p < 0.08$ and therefore not significant) can explain the significant difference between senior economists, senior humanists, and senior police cadets ($p < 0.001$). However, by looking at the results reported by Ahmed, we see that the 60 humanities students chose stag significantly more often than the 60 economics students ($p = 0.0074$, χ^2-test). Thus, his results support the hypothesis that economics students act more in accordance with

the game theoretical recommendations compared to humanities students. The result stated in table 3.13 is the result from my analysis.

Finally, Haucap and Müller [2014] conducted an experiment with 577 participants using a trust game to examine the behaviour of economics students and law students respectively. Starting with the decision of player 1 to trust or distrust, Haucap and Müller [2014, p.6, p.17 appendix A, figure A.1] found that economics students are more likely to distrust compared to law students ($p = 0.004$). Regarding player 2's decision on whether to reciprocate or defect, economics students were also significantly less likely to reciprocate compared to law students with $p = 0.006$ [Haucap and Müller, 2014, p.7, p.17 appendix A, figure A.2]. Haucap and Müller [2014, p.1, abstract] argue that the difference observed between economists and non-economists is largely due to a gender effect. However, the gender effects reported are smaller than the education effects, so it seems doubtful that they can suffice as an explanation for the differences. Haucap and Müller [2014] do not provide the gender distribution for the different subgroups (law/economics students and junior/senior). Thus, it is not possible to reconstruct their analysis or reported results.

Summarising the experimental results using social dilemmas, we see a difference in behaviour between economists and non-economists for public goods games and for prisoner's dilemma when participants are not allowed to make promises before playing. In situations where participants are allowed to make promises, the difference disappears [Frank et al., 1993] or is even reversed [Hu and Liu, 2003]. Finally, we also see a behavioural difference between economists and non-economists when playing stag hunt and trust game. Together with the findings from zero-sum games, the results from classroom experiments show that there is a difference in behaviour between economists and non-economists with economists acting more in accordance with the endorsements of textbook microeconomics.

3.3 Survey experiments

The second way scientists have tested whether there is a behavioural difference between economists and non-economists is via surveys. The focus of these survey can broadly be described under the four headings i) profit maximisation, ii) price raises, iii) resource allocations, and iv) proneness to corruption. The four topics are concerned with questions that are to some extent considered in microeconomics. For each topic, I will describe the questions under consideration, the action or choice endorsed by textbook microeconomics, and the results of the surveys. I will consider each of the four topics in turn. The results are summarised in tables 3.14, 3.15, 3.16, and 3.17.

3.3.1 *Profit maximisation*

Rubinstein [2006] conducted a survey among economics, philosophy, law, and MBA students to see how the different groups will make a trade-off between profit maximisation in a company and laying off people in that company. The trade-off was presented in a table stating that laying off 26 employees will increase profit with 0.6 NIS millions. Laying off additionally 26 employees will further increase profit with 0.6 NIS millions. Finally, laying off additionally 44 employees (so that a total of 96 out of 196 employees will be laid off) will maximise profit by increasing it with 0.4 NIS millions. Laying off more than 96 employees will make profit decrease again.

Referring to the microeconomic *theory of the firm*, the aim of a company is always to maximise its profit, since the owners of the company, being rational individuals, will seek to maximise their income [Jehle and Reny, 2011, p.125]. Thus, textbook microeconomics endorses the decision to profit maximise and thus lay off the number of people that will achieve this goal.

Looking at the results provided by Rubinstein [2006, p.3], we see that 46.8% of the economics students chose to profit maximise. This is compared to 25.3% of the students who did not study economics. Using the data presented in table Q1, I conducted a Welch T-test showing that the difference

Study	Setting	Economic endorsement	Non-economists	Economists	Difference	Hypothesis Supported
Rubinstein [2006]	Profit maximisation vs. laying off employees.	Profit maximise.	25.3% choose to profit maximise.	46.9% choose to profit maximise.	$p < 0.001$***	**Yes**
Cipriani et al. [2009]	Profit maximisation vs. laying off employees.	Profit maximise.	38.2% choose to profit maximise.	46.9% choose to profit maximise.	$p < 0.001$***	**Yes**
Brosig et al. [2010]	Profit maximisation vs. laying off employees.	Profit maximise.	15% choose to profit maximise.	38% choose to profit maximise.	$p < 0.001$***	**Yes**

Table 3.14: **Summary of empirical findings for differences in choices regarding profit maximisation.** *** indicates a significance level of 0.001.

is significant with $p < 0.001$ (two-tailed test). The result is summarised in table 3.14. Rubinstein also asked readers from *Globes* to answer the same questionnaire. The results from this survey show the same difference between working economists and non-economists [Rubinstein, 2006, p.6].

Finally, Rubinstein tested the consequences of providing participants with a function, $p(x) = 2\sqrt{x} - 0.1x - 8$, for profit maximisation rather than a table. Here x is the number of employees and the result, $p(x)$, is the profit of the company in NIS millions [Rubinstein, 2006, pp.2-4]. When students were given a function, the difference between economics students and MBA and mathematics students disappeared, with 73%-77% of participants choosing to profit maximise.[13] From this, Rubinstein [2006, p.9] concludes that:

> This appears to support the intuition that presenting a problem mathematically, as we often do in economics, conceals the real-life complexity of the situation.

Cipriani et al. [2009] also conducted a survey using the same question as Rubinstein. Based on the survey results provided, we see that economics students choose to profit maximise 46.9% of the time while other students (law, tourism, language, and mathematics) on average choose profit maximisation 38.2% of the time. Using a Welch T-test, the difference is significant with $p < 0.001$ [Cipriani et al., 2009, p.460 table 3 and p.462 table 6].

[13] Philosophy students and law students were not asked to use a function due to their limited familiarity with mathematics.

Study	Setting	Economic endorsement	Non-economists	Economists	Difference	Hypothesis Supported
Frey et al. [1993]	Fair to raise prices for water on a hot day?	Price raise is fair.	15% think price raise is fair.	34% think price raise is fair.	$p < 0.001$***	**Yes**
Haucap and Just [2010]	Fair to raise prices for water on a hot day?	Price raise is fair.	22% think price raise is fair.	60% think price raise is fair.	$p = 0.001$***	**Yes**

Table 3.15: **Summary of empirical findings for differences in choices regarding the fairness of raising the prices of water bottles on warm days.** *** indicates a significance level of 0.001.

Finally, Brosig et al. [2010] also replicated the results from Rubinstein. They found that 38% of economics students choose to profit maximise while the same was true for 15% of the non-economics students. Again, the difference in choices is reported as significant with $p < 0.001$ (Two-sided Mann-Whitney U-test) [Brosig et al., 2010, p.49].

3.3.2 Price raises

Another survey was conducted by Frey et al. [1993] to test whether people find a price raise on water bottles on a hot day fair.

Considering the *general equilibrium* theory presented in microeconomics textbooks, a market system of goods is said to be in equilibrium when the prices of the goods are represented by a price vector p^*, such that there is no excess demand [Arrow and Debreu, 1954, McKenzie, 1954, Jehle and Reny, 2011, p.206]. In other words, in a full competitive market with rational agents, prices will adjust so that supply equals demand. If demand for water bottles increases due to a warm day, there will be an excess demand compared to the supply. However, by increasing the price of water bottles, the demand will decrease (since some people will no longer want to buy the water at the new increased price). Thus, a new market equilibrium can be reached by increasing the price of water bottles.

The results from Frey et al. [1993] show that 34% of economics students

consider a price raise fair compared to 15% of households in the same areas. The difference is significant with $p < 0.001$.[14] They further considered what would happen, if survey takers were told that the price raise was due to an excess demand for water bottles. Here, 52% of the economics students said the price raise was fair, while the number was 22% for households in the same areas. The difference remains significant [Frey et al., 1993, p.274, p.275 table 2]. Finally, the question was changed to a price increase in snow shovels after a snow storm. Here, 60% of economics students and 18% of households thought the price raise was fair, which is also a significant difference ($p < 0.001$).[15]

The study conducted by Frey et al. [1993] was replicated by Haucap and Just [2010] among students on a military university. They find the same difference between economics majors and non-economics majors and their results are robust when changing the framing of the questions [Haucap and Just, 2010, p.245].

3.3.3 Resource allocation

Faravelli [2007] used a survey to ask about fair allocations of resources (plants) between two individuals. There were three possible allocations [Faravelli, 2007, p.1407]. The first - following a *utilitarian* resource allocation principle - would maximise the total number of fruits produced, but at the expense of a very skewed allocation of the fruits. The second - following a *Rawlsian* resource allocation principle - had a lower total production of fruits and a somewhat equal distribution of the fruits. The final allocation - following an *egalitarian* resource allocation principle - had the lowest total number of fruits produced but with a completely equal distribution of the fruits. The question asking participants to choose between the three allocations of plants had four variations. In the first variant, no context was given. In the second, it was stated that the reason one individual produced more fruit than the other was that the latter had a disability. In the third variation, the unequal production

[14] Using the data reported in Frey et al. [1993, p.272 table 1] to perform a Welch T-test.
[15] Using the data reported in Frey et al. [1993, p.275 table 3] to perform a Welch T-test.

was explained by different work efforts. In the final variation, no context was given except for the minimum number of fruits needed for survival.

Questions of how to distribute resources among people are considered in microeconomics as *welfare economics*. Looking at a selected number of microeconomics textbooks, there is often no explicit recommendation concerning the different allocation methods. Varian [2014, ch.33, pp.631-642] introduces the three different allocation principles (utilitarian, Rawlsian/maximin, and egalitarianism), but does not endorse any of them. Instead, he argues that an equilibrium reached via competitive market mechanisms from an equal distribution of resources will be a fair (and efficient) allocation.

Jehle and Reny [2011, pp.282-290] also abstain from making a direct judgement in the choice between a Rawlsian and a utilitarian allocation principle:

> Once again, your choice of social welfare function is a choice of distributional values and, therefore, a choice of ethical system. The choice is yours [Jehle and Reny, 2011, p.290].

This said, the utilitarian allocation principle is introduced with the following sentence:

> The utilitarian form is by far the most common and widely applied social welfare function in economics [Jehle and Reny, 2011, p.284].

Further, pp.288-290 is devoted to a debate between Harsanyi - arguing "remarkably straightforward" for a utilitarian allocation principle - and Rawls, whose argument is not "wholly persuasive" and where

> there is little obvious justification for adopting such a decision rule, unless, of course, you are extremely (irrationally?) pessimistic [Jehle and Reny, 2011, p.290].

Thus, indirectly at least, Jehle and Reny [2011] seem to support the utilitarian allocation principle.[16]

Finally, Mas-Colell et al. [1995, pp.825-831] introduce the three different allocation principles, without making a judgement between them. Thus,

[16] This is an illustration of how economists can end up slipping in normative claims in their textbooks. I will discuss this in greater length in chapter 5.

Study	Setting	Economic endorsement	Non-economists	Economists	Difference	Hypothesis Supported
Faravelli [2007]	Choose between utilitarian, Rawlsian, or egalitarian resource allocation.	Choose utilitarian resource allocation.	16% choose the utilitarian resource allocation.	27% choose the utilitarian resource allocation.	$p = 0.0004$***	**Yes**

Table 3.16: **Summary of empirical findings for differences in choices regarding resource allocation between two individuals.** *** indicates a significance level of 0.001.

microeconomics textbooks in general seem careful not to endorse one principle (this being a normative judgement), however, there also seems to be some preference for the utilitarian allocation principle [Jehle and Reny, 2011].[17]

Going back to the results from the survey, Faravelli [2007, p.1409], found that economics students are more likely to endorse the utilitarian allocation principle compared to non-economics students when no context is given (p=0.0004). He also found that when the unequal utilities were described as due to a disability for one of the individuals, the majority of both economics and non-economics students chose the Rawlsian allocation principle [Faravelli, 2007, p.1414].

3.3.4 Proneness to corruption

The final surveys I will consider investigate the dilemmas between individual gain and honesty. Though microeconomics textbooks are not concerned with the dilemma between honesty and individual gain, some variants of the rationality assumption - for example found in consumer theory - will endorse that individuals maximise their own monetary gains. Using this as a guide for behaviour might suggest that individuals should seek to maximise their own gains regardless of how it affects others. The results from the four surveys considered on this topic are summarised in table 3.17.

[17] It should be noted that welfare economics is not necessarily a part of the microeconomics training that students receive.

Study	Setting	Economic endorsement	Non-economists	Economists	Difference	Hypothesis Supported
Frank and Schulze [2000]	Choice between accepting bribe or not.	Accept bribe.	No information provided.	No information provided.	$p < 0.01$**	**Yes**
Frank et al. [1993]	Choice regarding: 1) Reporting an additional computer. 2) Returning an envelope.	Do neither.	1) 23.3% would not report the computer. 2) 10.0% would not return the envelope.	1) 41.7% would not report the computer. 2) 29.2% would not return the envelope.	?	**Yes**
Laband and Beil [1999]	Choice regarding reporting the correct income to calculate dues.	The right income will not be reported.	Sociology association: 78% of dues collected. Political science association: 91% of dues collected.	Economics association: 93% of dues collected.	?	**No**
Yezer et al. [1996]	Choice regarding: 1) Reporting an additional computer. 2) Returning an envelope.	Do neither.	Biology: 1) 59.0% would not report the computer. 2) 72.4% would not return the envelope. Psychology: 1) 46.0% would not report the computer. 2) 65.7% would not return the envelope.	1) 56.9% would not report the computer. 2) 68.1% would not return the envelope.	$p > 0.1$	**No**

Table 3.17: **Summary of empirical findings for differences in choices regarding proneness to corruption.** ** indicates a significance level of 0.01.

Frank and Schulze [2000] conducted a survey asking students to choose a company to do a plumbing task for a university film club. Students were given a price list of the different companies along with a list of how much each company would bribe the student if they selected that company. Numbers correlate such that the more expensive the plumber is, the more the company will pay the student to choose them. Students knew that choosing a more expensive plumber would result in a greater amount of money for themselves, and less for the film club. The survey results show that economics students are significantly more likely to choose a more expensive company in order to get paid a greater amount themselves ($p < 0.01$) [Frank and Schulze, 2000, pp.106-107]. Frank and Schulze [2000] interpreted this as a sign that economics students are more corrupt than other students.

Similarly, Frank et al. [1993, pp.168-170] conducted - what they call - an honesty survey among economics and astronomy students. First, students

were asked to consider a small business owner receiving ten computers but only being billed for nine. Students were asked to indicated whether they believed that the owner would report the mistake, and whether they would report the mistake if they were the owner. Second, students were asked to imagine a person finding an envelope containing $100 and bearing the owner's name and address. Again, students were asked to indicate whether they believed the envelope would be returned and whether they would return the envelope, had they found it. For the economics students being taught traditional microeconomics, 41.7% would not report the additional computer and 29.2% would keep the envelope. For astronomy students the numbers were 23.3% and 10.0% respectively [Frank et al., 1993, p.169, figure 3]. No statistical test is reported and there is not sufficient information to make one.

Potentially contrary evidence is found by Laband and Beil [1999] who conducted a survey on income for members in economics, political science, and sociology associations. All three associations have income-based dues, and so Laband and Beil [1999] were able to calculate whether the income reported for due payments had the same distribution as incomes reported in the survey. Results show that - across the three disciplines - the amount of cheating (by reporting a smaller income for dues than one actually has) is proportional to how much one will gain by cheating. Economists could gain the least by cheating and so cheated less than both political scientists and sociologists [Laband and Beil, 1999, p.97].

Finally, Yezer et al. [1996] replicated the survey made by Frank et al 1993 and did not find the same differences between economics students and biology and sociology students ($p > 0.1$) [Yezer et al., 1996, pp.182-183].[18]

Summarising the survey results concerning differences in behaviour between economists and non-economists, we see a significant difference for dilemmas concerning profit maximisation, price raises, and resource allocation. When it comes to choosing personal gain rather than honesty, the picture is mixed. It should, however, be noted that the ethical trade-off between

[18]It should be noted that Frank et al. [1996] argue that Yezer et al. [1996] consider class averages, where Frank et al. [1993] look at individual differences.

honesty and personal gain is not discussed in microeconomics. Thus, the survey results show that there is a difference in responses between economists and non-economists for topics debated in microeconomics textbooks. Further, they suggest that providing participants with additional information about the situations can reduce the difference [Faravelli, 2007, Cipriani et al., 2009].

3.4 Observational studies

The final category I will consider is observational studies. This group contains experiments where data are reported from real life choices of economists and non-economists. The section is divided into two parts. First, I report the findings from the observational studies concerning differences in behaviour. Second, I discuss whether the empirical findings considered in this chapter support the hypothesis that microeconomics is self-fulfilling.

3.4.1 Envelopes and donations

To my knowledge, one contextualized experiment and two observational studies have been conducted to examine the difference in behaviour between economists and non-economists. I will discuss each of these in turn. The results from the studies are summarised in table 3.18.

Study	Setting	Economic endorsement	Non-economists	Economists	Difference	Hypothesis Supported
Yezer et al. [1996]	Do students return envelopes with $10?	Do not return envelopes.	31% of envelopes are returned.	56% of envelopes are returned.	$p < 0.1$	**No**
Frey and Meier [2003]	Do students donated to a fund supporting other students' tuition fee?	Do not contribute.	68.65% contributes to a fund.	61.80% contributes to a fund.	$p < 0.01$**	**Yes**
Bauman and Rose [2011]	Do students donate to fund 1 or fund 2?	Do not donate to either fund.	Fund 1: 8% donate. Fund 2: 14% donate.	Fund 1: 5% donate. Fund 2: 10% donate.	Fund 1: $p < 0.01$** Fund 2: $p < 0.01$**	**Yes**

Table 3.18: **Summary of empirical findings from contextualised experiments and observational studies regarding differences in behaviour and choices between economists and non-economists.** ** indicates a significance level of 0.01.

Yezer et al. [1996] also conducted a contextualised envelope experiment in order to consider whether economics students behave differently than other students.[19] In the experiment, they placed an envelope with $10 in a classroom before either economics classes or other classes. The envelope was not sealed and besides the money, it contained a note addressed to a person (that did not exist) saying that the money was payment for a loan made by that person. Each envelope was tracked via the name it contained. The aim of the experiment was to test which envelopes would be returned.

The situation created in the experiment tests whether students will choose a personal gain despite its potential costs for other people. As mentioned above, this is not discussed in microeconomics textbooks. However, one may take some variants of the rationality assumption to endorse maximising individual monetary gain regardless of how it affects others. If people use these variants of the assumption as a guide for behaviour, we would expect them not to return the envelopes.

The results from the envelope experiment show that 56% of envelopes placed in classrooms prior to economics classes were returned. In comparison, only 31% of envelopes placed in classrooms prior to other classes were returned [Yezer et al., 1996, p.181]. Further, Yezer et al. [1996] found some evidence that people returning envelopes placed prior to economics classes had made a greater effort tracking down the person, for whom the money was purportedly intended.[20]

Frey and Meier [2003] conducted an observational study investigating how much students donate to finance other students' studies at their university. Using a variant of the rationality assumption discussed above, textbook microeconomic theory would suggest that no one will contribute to a fund. The results are built on the choices of 28,568 students. Of the economics students, 61.8% contributed to a fund. The percentage was 68.65% for non-

[19] Although Yezer et al. [1996] refer to their envelope experiment as a natural experiment, it does not seem to be one since their experiment is carefully set-up. Still, the experiment is conducted in a less controlled environment compared to a typical laboratory experiment. Thus, I have chosen to refer to the envelope experiment as a contextualized experiment rather than an observational study.

[20] For a critical remark see Frank et al. [1996].

economics students [Frey and Meier, 2003, p.451, table 1]. The difference is significant with $p < 0.01$ [Frey and Meier, 2003, p.453, table 2].

Finally, Bauman and Rose [2011] also reviewed data on how likely different student groups are to donate to social programmes at their university. The data set contains 65,044 observations based on 8743 individuals [Bauman and Rose, 2011, p.319]. The students could choose to donate to two different funds. As before, textbook microeconomic theory can be interpreted as endorsing not contributing to either fund. The percentage of students donating to the first fund was about 5% for economics majors and 8% for non-economics majors. For the second fund, the percentages were 10% and 14% respectively [Bauman and Rose, 2011, p.322]. For both funds, the difference is significant with $p < 0.01$ [Bauman and Rose, 2011, p.324, table 3].[21]

Summarising the results from the contextualized experiment and observational studies, there is some evidence that economists are more prone to return envelopes compared to others. Further, the observational studies suggest that economists in general donate less compared to non-economists.

3.4.2 Discussion: do economists act differently?

Looking at the results from all experiments considering behavioural differences between economists and non-economists, the vast majority show that there is a difference in behaviour such that economists act more in accordance with the endorsements of microeconomics textbooks in situations considered by the textbooks. Further, it is worth noting that all papers suggesting that there is no difference between economists and non-economists are published prior to 2009. Thus, all papers published in the last decade support the hypothesis that there is a behavioural difference between the two groups.[22]

[21] That economists donate less is also supported by Frank et al. [1993] who conducted a survey showing that economics professor donate less to charity compared to professors of other disciplines. A difference in the amount donated to a fund is also found by Ifcher and Zarghamee [2018, pp.60,62].

[22] Besides students' educational background, gender [James et al., 2001, Haucap and Müller, 2014, Wang et al., 2011, Hu and Liu, 2003, Gerlach, 2017, Cappelen et al., 2015, Rubinstein, 2006], age [Lopes et al., 2015, Frank et al., 1993], political affiliations [Lopes et al., 2015, Haucap and Just, 2010, Frey and Meier, 2003], logical abilities [Carter and Irons, 1991],

On a note of caution, it is important not to make too generalized conclusions on the basis of the experimental results. First, we have seen that the context of the decision matters, and that the differences between economists and non-economists might be mitigated when more information is provided [Faravelli, 2007]. Second, the evidence span over a variety of situations discussed in microeconomics textbooks. Though we can conclude that economists act differently in these situations, we cannot conclude that economists act differently in *all* situations [Lanteri, 2008]. Third, as suggested by Ahmed [2008], the choice of reference group can have an impact on how different the behaviour of economists is. Finally, most experiments are conducted with students in western societies. Thus, one should be careful not to generalise from student groups in western societies to the general population in western and non-western societies without further arguments or evidence [Bianchi, 1998].

Despite the notes of caution, the experimental evidence clearly shows that there is a difference in behaviour between economics students and non-economics students when choosing in economic situations discussed by microeconomics textbooks. Further, the results show that economics students act more in accordance with the endorsement of microeconomics textbooks compared to non-economics students. This gives us some indication that microeconomics is a self-fulfilling science. In order to show this, however, we also need to argue that the difference in behaviour can be *caused* by a *dissemination* of textbook microeconomics and is - for example - not only due to a self-selection effect where people choose to study economics *if* they already act in a certain way. I turn to this second part of my argument in the next section.

and demographic background data [Cipriani et al., 2009] have in some cases been significant predictors of choices. However, the educational effects are present even when controlling for these effects.

3.5 From learning to doing: indications of why economists act differently

Given that economists act differently than non-economists in economic situations, it is natural to ask whether this is due to their education, or if there is a self-selection effect such that people who have specific behavioural preferences also are more likely to choose to study economics. Much of the older experimental literature on the behaviour of economists is concerned with this question. I call the first effect - that the difference in choice is due to the economic training people receive - a *learning effect*. I call the second effect - that people with certain preferences self-select into economics programmes - a *selection effect*. Several authors claim to only find a selection effect [Brosig et al., 2010, Gandal et al., 2005, Frey and Meier, 2003, Frank and Schulze, 2000, Carter and Irons, 1991], while others also or only find a learning effect [Ifcher and Zarghamee, 2018, Haucap and Müller, 2014, Molinsky et al., 2012, Bauman and Rose, 2011, Wang et al., 2011, Haucap and Just, 2010, Cipriani et al., 2009, Faravelli, 2007, Frank et al., 1996, 1993].[23]

The question is relevant, since microeconomics can only be said to be self-fulfilling if there exists a learning effect and not only a selection effect. The dichotomy between the two effects, however, is hard to maintain. First, it is difficult to determine whether a selection effect for university freshmen is truly a selection effect, or if some of the individuals choosing to study economics have in fact been exposed to economic theory previously - for example through previous training, information from media, or their parents. Cipriani et al. [2009, p.463], for example, report that the probability of students choosing to study economics increases with 10% if their father is a senior manager, a member of the profession, or an entrepreneur. Further, it might be difficult to say whether senior economics students act different from freshmen economics students because of a learning effect or a double selection effect (choosing to continue studying economics) [Lanteri, 2008]. Thus, trying to determine whether there is a learning effect, by testing differences in

[23] Notice, that all papers after 2010 do confirm that a learning effect exists.

behaviour and choices between senior and junior university students, may confuse the debate rather than clarifying it. I therefore propose that a better way to test whether economics training has an effect on people's choices is to directly test the effect of exposing people to economic theory or concepts. If we can see a direct effect of exposing economists or non-economists to microeconomic theory, then this shows that dissemination of microeconomic theory can cause behavioural changes, which in turn will confirm that microeconomics can be self-fulfilling. Here, I first consider indications of a learning effect found in the framing and priming literature. Second, I present a study conducted by Ifcher and Zarghamee [2018] that directly tests whether a lecture in microeconomics can affect people's choices.

3.5.1 Evidence from framing and priming experiments

The literature on *context framing* contains some studies considering how framing a game as a market situation or competitive situation rather than a cooperative situation can change people's behaviour.[24] First, several experiments have been conducted to test whether participants choose differently in a prisoner's dilemma game, if the game is called a *community game* or *the cooperation game* compared to a *Wall street game* or *stock market game*. Here, Kay and Ross [2003], Liberman et al. [2004], Ellingsen et al. [2012] found that changing the name can either make the participants cooperate or defect more, depending on the name.[25]

Second, Engel and Rand [2014] ran a study where they framed a one-shot prisoner's dilemma game by either telling participants that they were on a team with the other player or that they were competing with the other player in setting prices in a market. Comparing the behaviour to a control group where the game was described in a neutral language, Engel and Rand [2014, p.387] found that participants being told that they were in a market situation

[24] I define context framing as proposed by Gerlach and Jaeger [2016, p.3]: context framing shapes players' experience of the game by 1) associating the game with different entities and/or 2) stressing specific aspects of the game.

[25] Notice that Belaus et al. [2018] found no significant effects when trying to replicate the findings of Kay and Ross [2003].

defected significantly more than participants in the control group (χ^2-test with $p = 0.011$). Interestingly, there were no significant difference in the choices made by participants exposed to the cooperative framing compared to the control group (χ^2-test with $p = 0.815$). This indicates that people choose to act more in accordance with textbook microeconomics if they are told that they are in an economic context or market situation.

The literature on experiments conducted with *priming* also suggests that indications of an economics setting or context can change participants' choices:[26] Kay et al. [2004] show that priming people with objects associated with business - or pictures of such objects - will make them view a given situation as more competitive. Further, priming people with business related objects made them propose smaller offers in an ultimatum game, and made them cooperate less in a prisoner's dilemma game called *the situation* (though the effect disappeared if the game was called *the community game*).

Molinsky et al. [2012] finally considered how priming participants with economic concepts affects their compassion when delivering bad news. They found that compassion is reduced by priming participants with economic concepts, and that the reduction is caused both by reduced empathy for the other person and because it is perceived as unprofessional to show compassion in such situations. This, again, indicates that there might be a connection between economic theory and concepts and how people choose to behave in situations where they are perceived to apply.

3.5.2 Testing the effects of receiving a lecture in microeconomics

Finally, we can consider the direct effects of receiving a lecture in microeconomics. Ifcher and Zarghamee [2018] conducted an extensive experiment to

[26] Broadly speaking, priming refers to the activation of mental concepts through subtle stimuli or situational cues that can facilitate impressions, judgments, goals, and actions [Bargh and Chartrand, 2000, Molden, 2014, Cohn and Maréchal, 2016]. Here, I focus on two experiments concerning prime-to-behaviour effects via situation-perception. This type of priming is used to test how situational cues can change people's behaviour by affecting their perception of a situation [Smeesters et al., 2010].

test participants' choices before and after reading a short lecture in microeconomics. The study was conducted in three stages. First, participants were asked to play a series of games in random order.[27] They were not informed about the outcome of the games before proceeding to the second stage. Second, participants were asked to read a small lecture in microeconomics. Participants were randomly assigned to three different lectures. All three lectures were designed to be parallel in structure to and to mirror the words used in standard microeconomics textbooks [Ifcher and Zarghamee, 2018, p.57]. The first lecture related the standard microeconomic model of how rational agents will play a game. The second lecture informed participants of the average behaviour of people (with no education in economics) playing economic games in experiments. The third lecture described the different ways games can be represented in economic theory. The three lectures are referred to as *normative*, *positive*, and *control*, respectively. The normative lecture is presented below. All three lectures also included an application of the corresponding lecture to the prisoner's dilemma game (called the *Box Game*) and to the ultimatum game (called the *Offer Game*) [Ifcher and Zarghamee, 2018, p.57]:

> **How to play games such as those you just played.** Normative economics helps economists understand how individuals should make decisions in games such as those you just played. To make normative economics assertions, economists build economic models. In such models, economists make the following assumptions: (1) that all individuals are self-interested and (2) that all individuals attempt to maximize their payments. Further, economists examine all the strategies available to an individual to determine which one maximizes his or her payment. Economists do this by working backward. First, economists consider all the choices the individual's opponent could make, and then, determine the choice that maximizes the individual's payment. Now we will apply normative economic analysis to the Box Game and the Offer Game to see what we can learn.

In the third and final stage of the experiment, participants were asked to play the same games that they played in the first stage.

[27] I have already reported the results from these games in section 3.2.

Comparing the change in choices for participants before and after having received the normative lecture with the change in choices for participants before and after having received the control lecture, we see that participants receiving the normative lecture significantly decrease their offers in the ultimatum game ($p < 0.001$), significantly decrease their minimum accepted offers ($p = 0.016$), significantly increase their defection rate in prisoner's dilemma ($p = 0.015$ for participants who answered the comprehension questions for PD correctly), and significantly reduce the donations they make in the dictator game ($p = 0.005$). No significant effect was observed for the public goods game [Ifcher and Zarghamee, 2018, pp.58-59]. Interestingly, Ifcher and Zarghamee [2018] also found that within the subgroup of participants receiving the *normative* lecture, participants who had prior exposure to economic theory (i.e. had taken economics courses) were significantly more likely to act in accordance with economic theory after receiving the lecture compared to participants with no prior economics exposure [Ifcher and Zarghamee, 2018, p.61, p.63 table 13].[28] When comparing choices made by participants before and after the *positive* lecture with the choices made by participants before and after the *control* lecture, no significant differences were found [Ifcher and Zarghamee, 2018, p.59].

Thus, Ifcher and Zarghamee [2018] provide clear evidence that standard microeconomics teaching (as the text provided in the *normative* lecture) can affect people's behaviour such that they act more in accordance with the behaviour endorsed by microeconomics textbooks.

3.6 Conclusion

In this chapter, I have argued that dissemination of textbook microeconomics can cause people to change their behaviour to be more in accordance with microeconomic theory, making microeconomics a self-fulfilling science. I

[28] n=43 for participants with prior economic exposure and n=44 for participants with no prior exposure. $p = 0.027$ for UG minimum acceptance offer, $p = 0.006$ for increase in defection rate in PD, and $p = 0.050$ for decrease in DG divisions. There was no significant effect for UG offers or PGG.

made the argument in two steps. First, I conducted a comprehensive review of the experimental literature considering whether economists act more in accordance with the endorsed behaviour of microeconomic theory compared to non-economists. The results from sections 3.2, 3.3, and 3.4 clearly show that this is the case in economic situations. Second, I argued that the empirical literature shows that dissemination of microeconomic theory and concepts can cause this behavioural change. Thus, microeconomics can be used as an example of a self-fulfilling science, since dissemination of its textbook theories can change people's behaviour in economic situations. For non-economic situations, or for situations where there are salient non-economic priorities (such as promises or an inclination to support people with disabilities) the empirical literature shows that the self-fulfilling effects of textbook microeconomics decrease.

Having established that microeconomics is an example of a self-fulfilling science, we are now ready to consider *how* it becomes self-fulfilling. The first step in doing this, is to analyse how economics is related to the normative claim that it is beneficial to act in accordance with one's own interests. This is the topic of the next chapter, where I also address the argument made by Friedman [1953] that it is possible to distinguish between positive and normative economics and that the former is independent of the latter.

Chapter 4

The benefits of self-interest

4.1 Introduction

In 1953, Friedman wrote a now canonical paper on the methodology of economics [Mäki, 2009]. In the paper, he argues that we can distinguish between *positive* and *normative* economics. Positive economics is concerned with how the world *is* and aims at making predictions about the world [Friedman, 1953, pp.145,148]. In contrast, normative economics is concerned with how the world *ought to be* and is thus relevant for policy recommendations and legislation [Friedman, 1953, pp.145-146]. According to Friedman, normative economics depends on positive economics, but positive economics is independent of normative economics. Thus, Friedman argues that positive economics does not provide any normative judgements on how the world (or our behaviour) *ought* to be. This view has been adopted by most microeconomics textbooks, where only social choice theory and welfare economics is described as normative economics [Jehle and Reny, 2011, p.267].[1] Thus, most microeconomics models are claimed to be positive models that do not contain any normative judgements and which aim is to make predictions about real-world behaviour.

[1] See section 5.3.

If it is correct that positive economics is independent of normative judgements, and if microeconomics is primarily positive, one may dispute that there is a connection between microeconomics and the normative claim that it is *good* to act (narrowly) self-interested. The aim of this chapter is to show that the claim has been a part of modern economic theory since its beginning and that it is still present in contemporary neoclassical economics. By going back to the original arguments for the benefits of greed and self-interest and tracing their influence to contemporary microeconomics, it becomes clear that even though economists have claimed that there are no normative judgements in positive economics, the assumption that individuals always strive to maximise their own gains (however defined) is closely related to the normative idea that it is beneficial to promote greed and self-interested actions among individuals in a market. It also highlights the point that historically as well as in contemporary microeconomics, positive economics includes several normative judgments, such that the divide between positive and normative economics is not maintained in practice.

The chapter is divided into three sections. In section 4.2, I describe how greed and commerce came to be seen as beneficial in the 18th century western Europe and discuss the argument made by Mandeville in defence of greed. In section 4.3, I present and discuss Adam Smith's argument that self-interest can lead to publicly beneficial outcomes for society. Finally, in section 4.4, I turn to the arguments by Voltaire and Hume on how self-interest can foster individual virtues. In all three sections, we see how the normative claim that self-interested actions are beneficial has been and still is present in positive economics theorising.[2]

4.2 Promoting a lesser evil: why greed is good

The argument that greed is beneficial for society was first stated in the early 18th century. The underlying motivation for it can be traced back to political

[2]The idea for this chapter and the choice of historical sources for it is indebted to Alex Voorhoeve's lectures in *philosophy of economics*, PH211/PH413, which I was fortunate to be GTAing in LT 2014/2015.

theorists in the renaissance and their attempts to improve government within the existing order [Hirschman, 1997, pp.12-15].[3] Rather than considering what an ideal society should look like, they began to explore how best to govern real states [Machiavelli, 1532, ch.15]. In the 17th century, this approach was extended to the study of human nature. Political philosophers such as Spinoza, Hobbes, and Rousseau argued that in order to study human nature, we have to consider people as they really are. Though they differ in their accounts of humans, they all assert that humans are subject to their passions and that these cannot be controlled by reason alone.[4] By the 18th century this claim was generally accepted, and theorists began to consider how to control the wilder, *violent passions* by other *calmer passions* [Hirschman, 1997, pp.24-27]. Calmer passions are human aspirations where an element of reflection and calculation is used to determine how to obtain them [Hirschman, 1997, p.32]. One such passion is the pursuit of monetary gain or greed.

Looking at the debate at the time, at least four arguments were made to promote greed among people:

- First, promoting greed is beneficial since it can control the violent passions [Hirschman, 1997, pp.24-25,32]. This is further supported by the general perception - due to the then present aristocratic contempt for economics activities - that greed is harmless in its consequences [Hirschman, 1997, pp.56-59].[5]

- Second, it will create social stability via predictable human behaviour and good government [Hirschman, 1997, p.49]. It will create predictable

[3]Hirschman [1997] - first published in 1977 - is the first book to suggest a connection between the 18th century discussion of human nature and the normative foundation of classical economics [Hirschman, 1997, pp.ix, xxi and 138-143]. The argument presented in the book started an industry of academic writings that is still present today [Glazer, 1985, Klamer and Colander, 1990, Mansfield, 1995, Walzer, 2002, Force, 2003, Grant, 2008, Strange, 2015, Gudeman, 2016, Fukuyama, 2017, Granovetter, 2018, de Champs, 2019, Guiot-Isaac, 2019]. Given his influence, I will use his historical analysis as a starting point for my own investigation.

[4]See e.g. Hobbes [1651, ch.8, 15, and 17], Rousseau [1762, book 2, chapter 6], and Hirschman [1997, pp.13-14].

[5]The aristocracy did not have to engage in trade or other kinds of employment because they owned land and would typically be able to live off the rent from this land. For a later analysis of this phenomenon see Ricardo [1815].

human behaviour because greed is a constant passion, which is not directed towards specific people at specific times (like lust or anger may be) [Hume, 1738, book 3, part 2, section 2]. It will create good government by strengthening the position of what would come to be seen as the middle class [Hirschman, 1997, p.83]. This, according to Smith [1776, pp.251-253] will reduce the power of the noblemen which in turn will lead to the rule of law along with the liberty and security of individuals.

- Third, if a society is governed by the right institutions and laws, individual greed can lead to a flourishing industry and thus be beneficial for society overall [Mandeville, 1714].

- Finally, promoting greed can foster individual virtues - such as religious tolerance and justice - because greed can compel people to ignore their differences in market situations. These virtues can then diffuse into other parts of society [Voltaire, 1733, Hume, 1738].

In the remainder of the chapter, I will show and discuss how these arguments for the benefits of greed and self-interest have influenced economic theory and still do.[6] I start with the argument made by Mandeville that greed is beneficial for society.

4.2.1 Mandeville's argument: private vices and public benefits

In the poem *The grumbling hive* from 1705 and in his later book *The fable of the bees* from 1714, Bernard Mandeville argues that private vices can lead to public benefits. The argument has three parts.

First, Mandeville distinguishes between individual virtues and individual vices. According to Mandeville, an individual virtue is

[6] I order to limit the scope of the enquiry, I have chosen only to focus on the arguments given by Mandeville, Smith, Voltaire, and Hume. For a review of additional philosophers making similar claims, see Hirschman [1997].

> every Performance, by which Man, contrary to the impulses of Nature, should endeavour the Benefits of others, or the Conquest of his own Passions out of a Rational Ambition of being good [Mandeville, 1714, pp.48-49].

With this definition, Mandeville combines the two concepts of virtue used at the time. The first concept, originating within theology, states that virtue is "a transcending of the demands of corrupt human nature, a conquest of self, to be achieved by divine grace" [Mandeville, 1714, p.xlvii]. While the second concept of virtue states that virtue is "conduct in accord with the dictates of sheer reason" [Mandeville, 1714, p.xlvii]. By combining the two concepts, Mandeville stresses that only actions that are in complete denial of one's nature by being both unselfish and dispassionate will count as virtuous actions.

In contrast, Mandeville defines individual vices as

> every thing, which, without Regard to the Publick, Man should commit to gratify any of Appetites [...] if in the Action there cou'd be observed the least prospect, that it might either be injurious to any of the Society, or even render himself less serviceable to others [Mandeville, 1714, p.48].

Thus, any action that is not virtuous is - according to Mandeville - vicious. When considering the actions performed by humans, Mandeville finds no action that is entirely done out of dispassionate selflessness. Thus, Mandeville argues that all actions undertaken by humans are vicious.

Second, Mandeville considers the public outcomes of individual actions. Mandeville defines public benefits as that which is useful or productive for national prosperity or happiness [Mandeville, 1714, pp.xlviii-xlix]. Using this standard to evaluate outcomes of - for example - trade and industry, Mandeville argues that they are beneficial for society. Since all actions at an individual level are vicious, Mandeville concludes that vicious individual actions can be beneficial for society.

Finally, Mandeville considers whether virtuous actions can have the same beneficial outcomes for society. If everyone lives virtuously, all will be content with a quiet life and no one will increase the wealth of society by furthering

industry, consuming luxuries, or trying to distinguish themselves from others. Thus, while a society filled with virtuous people might be preferable, it will not lead to the same amount of prosperity as a society with vicious people governed by law.

The three parts of Mandeville's argument are all present in his satirist poem about a grumbling bee hive [Mandeville, 1714, pp.17-37]:

> A Spacious Hive well stockt with Bees,
> That liv'd in Luxury and Ease;
> ...
> The worst of all the Multitude
> Did something for the Common Good.
> ...
> Thus Vice nurs'd Ingenuity,
> Which join'd with Time and Industry,
> Had carry'd Life's Conveniencies,
> It's real Pleasures, Comforts, Ease,
> To such a Height, the very Poor
> Liv'd better than the Rich before,
> And nothing could be added more.
> ...
> *Fraud, Luxury and Pride must live,*
> *While we the Benefits receive:*
> ...
> *So Vice is beneficial found,*
> *When it's by Justice lopt and bound;*
> ...
> *Bare Virtue can't make Nations live*
> *In Splendor; they, that would revive*
> *A Golden Age, must be as free,*
> *For Acorns, as for Honesty.*

Two things should be noted with regard to Mandeville's argument. First, though he is often considered one of the first proponents of a *laissez-faire* market economy (with no governmental interference in the market) [Mandeville, 1714, pp.xcviii,lx], Mandeville stresses that the public benefits only occur insofar as people's vices are checked in a market system and governed by a law that punishes harmful and criminal vices:

> When I assert, that Vices are inseparable from great and potent Societies, and that it is impossible their Wealth and Grandeur should subsist without, I do not say that the particular Members of them who are guilty of any should not be continually reprov'd, or not be punish'd for them when they grow into Crimes [Mandeville, 1714, p.10].

Without a successful law to control people, it is in no way certain that private vices will lead to public benefits. Second, Mandeville does not believe that individuals will ever be able to change their vicious behaviour. Thus, Mandeville's point is that humans will be vicious regardless of what we do, but with the right control of law, this viciousness can be beneficial for society as a whole [Mandeville, 1714, p.I].

Discussion of Mandeville's argument

Mandeville's argument - that individual vices can lead to public benefits - depends on very specific and uncommon definitions of individual virtues and vices and on an outcome-oriented definition of public benefits.

Looking at Mandeville's definition of virtues, it is idiosyncratic and not commonly accepted in his time, since it does not allow *any* actions to be virtuous. This is contrary to the different definitions of virtue presented - for example - in Hume [1738], Smith [1759], or modern philosophy [Hursthouse and Pettigrove, 2018, section 1.1]. Using a more common definition of virtues would allow for some of our actions to be virtuous, and so the argument that all human actions are vicious will fail.

The reason Mandeville believes that all actions are vicious is that they can be *explained* by self-interested motives such as impulses and desires. The view that all human actions are based on self-interested motives has been held by several philosophers throughout the time, including Hobbes [1651] and d'Holbach [1770, ch.11]. However, the problem with the view is that it infers from "all actions *can be explained* by self-interested motives" to "all actions *are performed* because of self-interested motives". While it is always possible to make a *post hoc* explanation - after an event has occurred - for

why an action could be motivated by self-interest, it does not follow that it actually was motivated by self-interest.

Using contemporary philosophy of science to evaluate the claim that all human actions are selfish (because they can always be explained as such), we see that the claim cannot be falsified by any empirical observation. This, according to Popper, makes it pseudo-scientific [Popper, 1963]. Another problem raised by Popper [1963] is that the claim cannot be used to *predict* human behaviour. Consider the example of a child who is drowning. An adult will be able to rescue the child by sacrificing their own life. Assume that the adult does not rescue the child. This action can be explained with their self-interest in staying alive. However, assume that they do rescue the child. This action can also be explained by their self-interested aim for praise and a good reputation. Thus, regardless of what the adult chooses, it can be explained by self-interest, and their choice can therefore not be predicted from the claim that they will act in their own interest.

Finally, Mandeville's argument that individual vices lead to public benefits depends on using two different measure of goodness when evaluating public outcome and individual actions. If Mandeville used the same measure of good consequences for both individual actions and public benefits, several individual actions will be beneficial rather than vicious.

4.2.2 The legacy of Mandeville's argument

Mandeville is one of the first persons to voice the normative claim that individual vices can be beneficial for society. I end this section by showing how his argument has influenced economists in the past three centuries and is still present in microeconomics today.

According to Horne [1981, p.559], Mandeville's argument became a silent reference point for much of the social thought and discussion in the 18th century. This influence is evident in for example Hume [1738, section 5, part 1] and Smith [1759, part 7, section 2, chapter 4] where both writers discuss him explicitly. Further, we can see Mandeville as a silent reference point when Smith writes:

> How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it [Smith, 1759, section 1, chapter 1, p.13].

Thus, Mandeville's argument was present in the minds of the early classical economists.

The legacy of Mandeville's argument can also be seen in later economists' writings. In *The general theory of employment, interest and money*, Keynes [1936, ch.23, part 7] cites Mandeville as an example of an early proponent of his ideas. Presenting an excerpt of Mandeville's poem, and quoting two passages from Mandeville's commentaries [Mandeville, 1714, vol.1, (Q.)B, pp. 199,215], Keynes uses Mandeville to defend his own macroeconomic position: that individuals should spend (rather than save) in order to boost the economy, and that the government should increase its public spending, in order to increase demand for various goods and services, leading to a decrease in unemployment.[7]

Another influential economist, who used Mandeville as a historical support for his economic theory is Hayek. Hayek was a member of the Austrian school of economics. He is known as one of the founders of the Mont Pèlerin Society, and was a prominent figure in the establishment of the Chicago School of economics.[8] While at the University of Chicago, Hayek delivered a lecture called *Dr. Bernard Mandeville. Lecture on a master mind* [Hayek, 1966]. According to Hayek, Mandeville was the first thinker to lay the foundation for the idea of a spontaneous growth of orderly social

[7] For further discussion of Keynes's reading of Mandeville see Lagueux [1998].

[8] The Chicago School of economics originated at the University of Chicago where some of its founders - such as Knight, Viner, and Simons - worked in the 1930s [Miller Jr, 1962, p.64]. In the 1940s, a conscious effort was made to create an economics department with prominent economists - such as Hayek and Friedman - who would advocate a private-enterprise economy with free market and limited government [Miller Jr, 1962, p.65] and [Van Horn and Mirowski, 2009, p.9]. This ideology was seen as essential for countering totalitarian societies and as a way to reduce government intervention in the affairs of corporations. (The latter being the goal of Luhnnow who financed the school though the Volker Fund from 1945 [Van Horn and Mirowski, 2009, pp.2,15,21,23,24].) For further discussion of Hayek's influence on the Chicago School see Caldwell [2011] and Van Horn [2015].

structure [Petsoulas, 2013, p.6] which was taken up and developed by Montesquieu, Hume, Tucker, Ferguson, Smith, and finally Menger [Hayek, 1967, p.99].[9] Hayek uses these earlier thinkers to support his own theory of spontaneous order [Whyte, 2019, p.162], where self-interested human activity at an individual level unintendedly creates an order that has a structure and is *beneficial* for the individuals involved [Petsoulas, 2013, p.12].[10] According to Hayek, the theory of spontaneous order shows that a free competitive market and limited government is the best political system for society. Further, he uses it to claim that earlier economic crises can be explained by people's attempts to *intervene* via governmental planning. Thus, Hayek - like Keynes - uses Mandeville as an authority to give historical weight to his own economic theory.

Finally, Mandeville's argument can be related to the 20th century claim that "greed is good" [Wight, 2018, p.6]. While Mandeville does not say that greed is good, it is not an uncommon thought in neoclassical economics that greed in fact *is* good because it results in public benefits. This can for example be seen in the introduction to Arrow and Hahn's economics textbook *General competitive analysis*, where they write:

> The immediate "common sense" answer to the question "What will an economy motivated by individual greed and controlled by a very large number of different agents look like?" is probably: There will be chaos. That quite a different answer has long been claimed true and has indeed permeated the economic thinking of a large number of people who are in no way economists is itself sufficient grounds for investigating it seriously [Arrow and Hahn, 1971, p.vii].

As noted in the quote, the sentiment that greed is good has also diffused from economics into business and popular culture, which can for example be seen in the speech made by Gordon Gekko in the 1987 movie *Wall street*. Here, Gekko, famously concludes his speech with the following words:

[9] I will return to Menger in subsection 5.2.3.

[10] According to Jacobs [2000], the concept of a spontaneous order was used in the 20th century by Polany prior to Hayek's use and adaption of the concept. However, the concept is now primarily associated with Hayek [Whyte, 2019, p.161].

> The point is, ladies and gentleman, that greed - for lack of a better word - is good.
> Greed is right.
> Greed works.
> Greed clarifies, cuts through, and captures the essence of the evolutionary spirit.
> Greed, in all of its forms - greed for life, for money, for love, knowledge - has marked the upward surge of mankind.
> And greed - you mark my words - will not only save Teldar Paper, but that other malfunctioning corporation called the USA [*Wall street* 1987, Gordon Gekko].

Though the phrase "greed is good" is often associated with Smith's theory "of an invisible hand" it is important to stress that Smith's argument is not related to the idea that greed is good [Otteson, 2002, Evensky, 2005, Young, 2009, McCloskey, 2010, Wight, 2018, pp.6-7]. The origin of "greed is good" is Mandeville.

Is greed good?

Looking at the later uses of Mandeville's argument, we see that both Keynes and Hayek use Mandeville to support their own theories. Since Keynes focuses on the macroeconomics rather than microeconomies [Lagueux, 1998, Wight, 2018], I will not consider his theory in further details.

Hayek uses Mandeville as a first source to his idea that a spontaneous order will occur in free markets with no governmental interference. I will discuss the merit of his theory in subsection 4.3.2. Here, it suffices to note that Hayek's argument is not supported by Mandeville, since the latter clearly states that the beneficial outcomes of commerce *only* occur under a rule of law which is more demanding than the use of law Hayek will accept.[11]

Finally, whether greed is good will depend on 1) the context where people act greedily, 2) the institutions governing that context, and 3) our measure of "goodness" used to determine whether greed is good in that context. According to Mandeville, greed is not good since it is a vice. Further, greedy

[11] For further discussion of Hayek's interpretation of Mandeville, see Petsoulas [2013, ch.3].

actions can only create good outcomes for society if they are checked by laws and institutions. Thus, even though the claim "greed is good" can be traced back to Mandeville, it is unlikely that Mandeville would accept it.

4.3 Why self-interested behaviour is beneficial for society

4.3.1 Smith's arguments in defence of self-interest

The normative claim that self-interested behaviour can lead to beneficial outcomes for society is also a theme in Smith's *Wealth of nations*. Here, Smith makes the claim at least two places.[12]

First, Smith [1776, book 3, chapter 4, pp.259-268] provides three reasons why increased commerce - caused by people's (narrow) self-interest - is beneficial for a country:

a. It will enable farmers to sell their goods and encourage them to increase their production.

b. The merchants who accumulate wealth in the cities will use it to buy uncultivated lands in the countryside since "Merchants are commonly ambitious of becoming country gentlemen" [Smith, 1776, p.259]. This is beneficial for the country because it will increase the cultivation of the countryside and thus the gross production of goods. Further, merchants are accustomed to make profit, which makes them better suited than noblemen to cultivate and improve the lands.

c. It will introduce order and good government in the country by reducing the number of internal wars since commerce makes it beneficial for people to trade rather than make war upon each other. This, in turn, will increase the liberty and security of people.[13]

[12] I will return to Smith and his account of human behaviour in subsection 5.2.1.

[13] Smith [1776, pp.251-253] makes a related argument for how commerce increases stability in a country.

Given the three reasons, Smith [1776, p.268] concludes that

> [i]t is thus that through the greater part of Europe the commerce and manufactures of cities, instead of being the effect, have been the cause and occasion of the improvement and cultivation of the country.

Evaluating Smith's three reasons, I see no problem in the assertion that increased commerce will give farmers a place to sell their goods and an incentive to produce more goods. The second claim, that city merchants will spend their money buying up land on the countryside in order to become country gentlemen, seems like a time specific observation that will not apply today. Finally, the idea that commerce can reduce war and increase good government within a country is an empirical claim that is outside the scope of this thesis.

Second, in book four, chapter two, Smith uses his famous *invisible hand* metaphor to make an argument for the public benefits of self-interest. The context of the argument is a critique of the mercantilist policy to monopolise colonies' imports and exports [Kennedy, 2009, p.251]. Here, Smith argues that merchants, who will not risk overseas trade, will try to buy the monopolised goods at home, mitigating the negative effects of the policy. Thus, Smith concludes that merchants seeking their own interest unintentionally help society by reducing the negative effects of mercantilist laws:

> every individual necessarily labours to render the annual revenue of the society as great as he can. He generally, indeed, neither intends to promote the public interest, nor knows how much he is promoting it. By preferring the support of domestick to that of foreign industry, he intends only his own security; and by directing that industry in such a manner as its produce may be of the greatest value, he intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. [...] By pursuing his own interest he frequently promotes that of the society more effectually than when he really intends to promote it. I have never known much good done by those who affected to trade for the publick good. It is an affectation, indeed, not very common

> among merchants, and very few words need be employed in dissuading them from it [Smith, 1776, pp.291-292].

Later, in the same context, Smith elaborates:[14]

> It is thus that the private interests and passions of individuals naturally dispose them to turn their stocks towards the employments which in ordinary cases are most advantageous to the society. But if from this natural preference they should turn too much of it towards those employments, the fall of profit in them and the rise of it in all others immediately dispose them to alter this faulty distribution. Without any intervention of law, therefore, the private interests and passions of men naturally lead them to divide and distribute the stock of every society among all the different employments carried on in it as nearly as possible in the proportion which is most agreeable to the interest of the whole society [Smith, 1776, book 4, chapter 7, part 3].

Here, Smith argues that if governments do not restrict which employments people can pursue, the gain and loss in profit for different trades will make sure that people change their employment in accordance with what is beneficial for society. In the next passage, however, we are again reminded that Smith writes this argument against mercantilist laws:

> All the different regulations of the mercantile system necessarily derange more or less this natural and most advantageous distribution of stock. But those which concern the trade to America and the East Indies derange it perhaps more than any other [...] Monopoly is the great engine of both [...] Monopoly of one kind or another, indeed, seems to be the sole engine of the mercantile system.

Thus, Smith's argument that self-interest in a market setting can lead to a publicly beneficial outcome when unregulated by the government is presented in a very specific context against mercantilist attempts to monopolise the market [Viner, 1927, p.210].

Given the context of Smith's argument, it does not support the claim that individual self-interest *in general* will lead to beneficial outcomes for society

[14] This specific quote has been edited out [p.365] of the 2008 edition otherwise used.

or that Smith is against *all* governmental interference in markets. The latter has - for example - been shown by Viner [1927, pp.217-231], who summarises the governmental projects and interventions that Smith approves of. These includes free public education and potentially health care, public building projects, legislation on interest rates, and taxation. Thus, Viner concludes that

> Adam Smith was not a doctrinaire advocate of laissez faire. He saw a wide and elastic range of activity for government, and he was prepared to extend it even farther if government, by improving its standards of competence, honesty, and public spirit, showed itself entitled to wider responsibilities [Viner, 1927, p.231].

4.3.2 The legacy of Smith's invisible hand metaphor

Economists - especially since the 1970s - have praised Smith's metaphor of the invisible hand as the most important concept in economics and even in all of the social sciences [Samuels, 2011, pp.10-13], [Kennedy, 2009, pp.239-241]. Before looking at the contemporary use of the invisible hand metaphor, however, I want to stress that the metaphor is not originally Smith's and that it is not essential for his economics writings [Kennedy, 2009, pp.243, 253-254].

The metaphor of an invisible hand has been used in several earlier sources that Smith would have been familiar with [Kennedy, 2009, pp.242-243]. These include the *Iliad*, poems from ancient Greece, early Christian literature, and Shakespeare's Macbeth [Samuels, 2011, pp.21-25]. That the invisible hand metaphor was not essential for Smith's writings can be seen in at least two ways. First, Smith only uses the metaphor three times throughout all his writings: once in *Wealth of nations*, once in *Moral sentiments*, and once in *History of Astronomy* [Kennedy, 2009]. If it had been an essential part of his economic theory, it is curious that he uses it so sparsely. Second, the metaphor is not mentioned in early comments on *Wealth of nations* like Pownall [1776] or by early bibliographers of Adam Smith such as Stewart et al. [1793] [Kennedy, 2009, p.240]. Further, the classical economists inspired by Smith did not use the metaphor in their works. Thus, it is not mentioned by Malthus [1798], Ricardo [1817], Mill [1848], or Marx [1887]. In addition, the

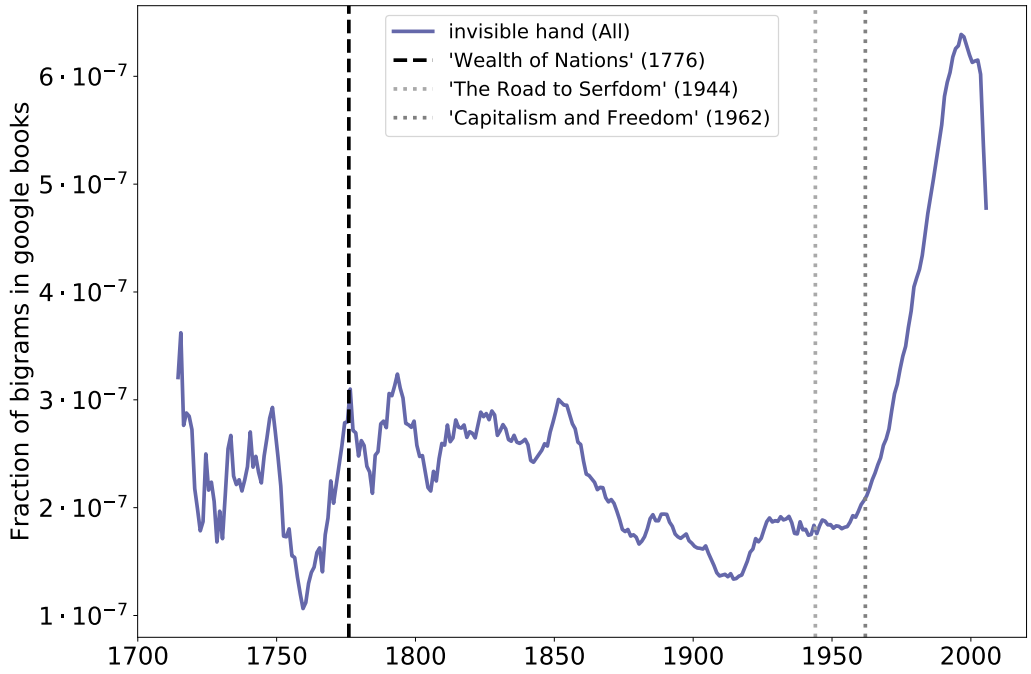


Figure 4.1: **Fraction of bigrams in google books samples being *invisible hand* between 1710 and 2010.** The line sums the use of the case-insensitive bigram *invisible hand* as a fraction of all bigrams in google books' book sample in the period between 1710 and 2010. The line is smoothened by taking a rolling average of 10 years. Source: Google Books Ngram Viewer, http://books.google.com/ngrams, [Michel et al., 2011].

paragraph containing the invisible hand metaphor was paraphrased without the metaphor by Buckle in 1857 [Kennedy, 2009, p.240]. Finally, looking at figure 4.1, we see that the use of *invisible hand* did not change dramatically due to Smith's publication in 1776. Taken together, this suggests that the status of the invisible hand metaphor in economics today is due not to Smith or his contemporaries but to its later uses and interpretations.

Turning to the current interpretations of the invisible hand metaphor, it has been used to promote both socialism and capitalism [Samuels, 2011, p.12]. However, the most famous and well-known interpretation of the metaphor is as part of an argument in support of a laissez-faire market economy [Samuels, 2011, pp.2,179,186]. This interpretation is likely to have originated

in Germany in the 19th century as a result of an intellectual hostility towards the British laissez-faire doctrine [Montes, 2008, p.159]. By the early 20th century, the interpretation was already well-known [Viner, 1927] and by the mid-20th century, it had become a standard part of microeconomics textbooks. This can for example be seen in Samuelson's widely used textbook from 1948:

> Even Adam Smith, the canny Scot whose monumental book, "The Wealth of Nations" (1776), represents the beginning of modern economics or political economy - even he was so thrilled by the recognition of an order in the economic system that he proclaimed the mystical principle of the "invisible hand": that each individual in pursuing only his own selfish good was led, as if by an invisible hand, to achieve the best good for all, so that any interference with free competition by government was almost certain to be injurious. This unguarded conclusion has done almost as much harm as good in the past century and a half, especially since too often it is all that some of our leading citizens remember, 30 years later, of their college course in economics [Samuelson, 1948a, p.36].

As already mentioned, however, Smith does not support a laissez-faire economy or even pay much attention to the idea of an invisible hand. Thus, the idea of an invisible hand in support of a laissez-faire economy seems to has taken on a life of its own independently of Smith and his argument. This is also suggested by figures 4.1 and 4.2, where we see that the use of "invisible hand" has gained popularity from the mid-20th century, and that the increase cannot be fully explained by a general increase in publications concerning Adam Smith.

Samuels [2008, p.185], Kennedy [2009], and Samuels [2011, ch.8] all argue that the use of the invisible hand metaphor from the 1940s onwards can be linked to the *ideological promotion* of capitalism and laissez-fair economy as a response to totalitarian regimes. This can, for example, be seen in the writings of Hayek, who - in his famous book *The road to serfdom* from 1944 - argues that the rise of Nazism in Germany was not due to Germans being "evil" but rather to preceding socialist policies and planning attempts. These policies had accustomed people's mindset to the idea that planning is good. When the

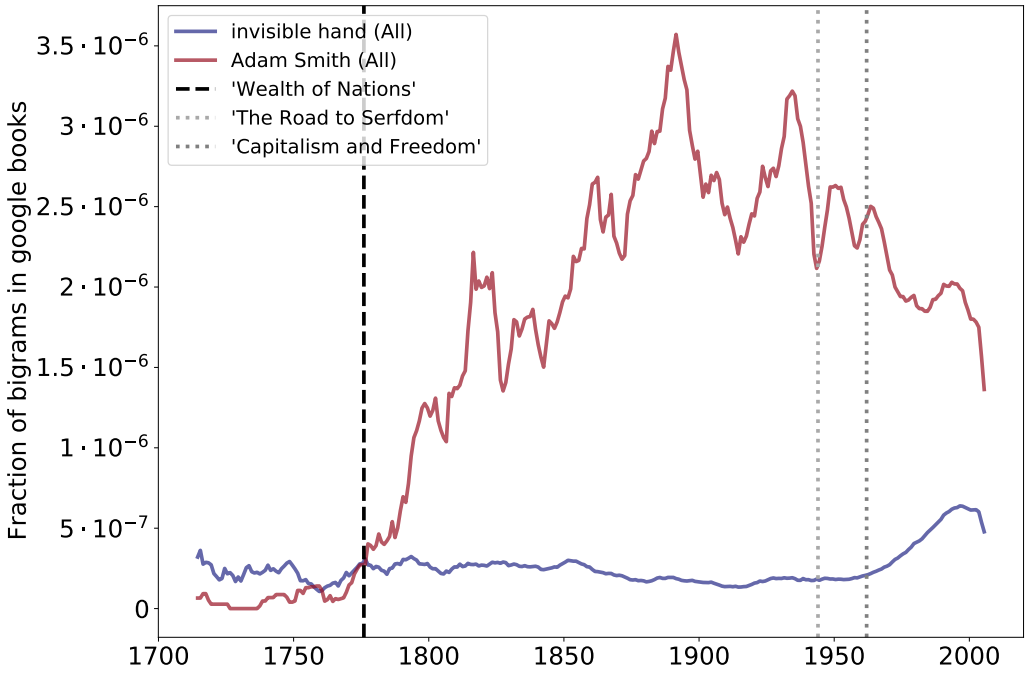


Figure 4.2: **Fraction of bigrams in google books samples being *invisible hand* and *Adam Smith* respectively between 1710 and 2010.** The two lines sums the case-insensitive use of each bigram. They are smoothened using a 10 year's rolling average. Source: Google Books Ngram Viewer, http://books.google.com/ngrams, [Michel et al., 2011].

politicians failed to make a successful plan, it led to confusion and mistrust in the public. This made it possible for a "strong man" to step forward and "recreate order" by introducing a totalitarian regime [Hayek, 1944, pp.11, 72-89]. Since Hayek believes that socialism was created and promoted in the intellectual circles rather than being a working-class phenomenon, he argues that the best way to counter it is to make reasoned arguments showing how a free market economy is the only possible foundation for real freedom [Hayek, 1944, pp.28,35], [Samuels, 2011, p.200]. In doing this, Hayek uses the invisible hand metaphor to argue that self-interested actions will lead to greater public benefits than altruistic actions:

> in fact we generally are doing most good by pursuing gain. [...] The aim for which the successful entrepreneur wants to use his

> profits may well be to provide a hospital or an art gallery for his home town. But quite apart from the question of what he wants to do with his profits after he has earned them, he is led to benefit more people by aiming at the largest gain than he could if he concentrated on the satisfaction of the needs of known persons. He is led by the invisible hand of the market to bring the succour of modern conveniences to the poorest homes he does not even know [Hayek, 1976, p.145].

As can be seen in figures 4.1 and 4.2, there has been a boom in literature mentioning the invisible hand from the 1960s. According to Samuels [2011, pp.19,201], there are two reasons for this increase. First, the concept has been more widely used since the establishment of mainstream neoclassical economics. Second, the concept has been used to promote a capitalist system during and after the cold war. Thus, the use of the invisible hand metaphor is not only descriptively but also ideologically motivated. This is also supported by Caldwell [2011], who writes:

> It is clearly true that free market ideology began to become more popular in the United States and Britain by the early 1980s, and worldwide after the collapse of the Soviet Union. It is also evident that individuals associated with both the Mont Pèlerin Society and the Chicago School are free market advocates, and that free market think tanks have grown rather dramatically in number worldwide, especially since the 1970s [Caldwell, 2011, p.311].

Finally, the legacy of the invisible hand metaphor is still present in microeconomics textbooks. Here, the first fundamental theorem of welfare economics is often called *the invisible hand theorem*. The theorem is an essential part of any microeconomics course and it states that a complete market with no transaction costs, where all agents have perfect information, where there is a price-taking behaviour, free exits and entries, no monopolies, and where it is always possible to find a new preferred bundle of goods arbitrarily close to an old one, will tend towards a competitive equilibrium that is Pareto optimal. The link to the invisible hand metaphor can, for example, be seen in Jehle and Reny [2011] where the theorem is accompanied by the following comment:

> [The theorem] provides some specific support for Adam Smith's contention that society's interests are served by an economic system where self-interested actions of individuals are mediated by impersonal markets [Jehle and Reny, 2011, p.217].

Is self-interest publicly beneficial?

Given the above discussion, it is clear that the invisible hand metaphor has been used in normative economics to promote capitalism and laissez-fair economy. However, we also see that the metaphor has been used in arguments presented as positive economics. Thus, it is relevant to ask whether there is empirical evidence that self-interested actions in a free market will always be publicly beneficial or whether this claim is in fact based on normative judgements used in positive economics.

As already mentioned, Hayek uses the principle of *spontaneous order* and the idea of unintended and unforeseen consequences to argue for the benefits of free markets [Samuels, 2011, p.201,203]. According to Hayek, a free market - controlled by the laws of property, tort, and contract - will produce socially beneficial outcomes from self-interested individual actions by creating the conditions under which any member of society can acquire and sell private property without violence or fraud [Jacobs, 2000, pp.54,56,128]. Trades made in the market will always be mutually beneficial since people can abstain from trading if they do not like the proposed exchanges. Thus, the market enables people with different aims, knowledge, and abilities to trade with each other, furthering their own interests - whether they are selfish or not [Hayek, 1976, p.110,113]. Finally, Hayek argues that markets will adapt to changes in demand, since these will lead to changes in prices, making some businesses more profitable than others. This, in turn, will make people change their occupations such that the markets can supply the goods people want [Hayek, 1976, pp.114,125].

Evaluating Hayek's argument for why self-interested actions in a free market is beneficial for society, it is first important to note that even if there is no violence or fraud in the market, people can still be exploited. Imagine, for example, a worker who has not eaten in a week. At this point, they might be

willing to work a full day or sell a kidney for a bowl of lentils. On Hayek's account, this is a mutually beneficial exchange, but one may wonder whether the worker would accept it, if they had had an alternative. Thus, it is not given that people can abstain from trading if they do not like the proposed exchange. This is especially important since it is nearly impossible to live outside a market system in today's society. Second, Hayek's arguments seem to indicate that *all* people will have the *same* opportunities to further their aims and exploit new opportunities. Unfortunately, this need not be the case. Depending on the socio-economic factors of their families, two equally gifted and motivated children can have very different opportunities in their lives. Finally, evidence from the research in advertisement and behavioural economics shows that people's preferences can be changed by different kinds of manipulations [Boyland and Halford, 2013, Sunstein, 2014, Dahlen et al., 2020]. Thus, especially larger companies may be able to influence people to want what they supply - even if they would not have wanted it, if they had not been exposed to manipulation.

Hayek is aware that in a free market setting, some people will lose their income or business and be unable to receive help. However, he believes that a free market will ultimately benefit everyone in it from the increase in aggregate supply of goods. Further, he argues that the observable harm happening to some people should not be used as an argument to prevent the diffused, unknown, and indiscriminate benefits of many [Whyte, 2019, p.164].[15]

Given Hayek's position, it is relevant to ask how we define *publicly beneficial outcomes* in positive economics. One definition is to say that any Pareto optimal distribution is publicly beneficial since everyone is made as well off as possible given their initial resources and the condition that no one is made worse off. If we use this definition, then the first theorem of welfare economics states that under some idealised circumstances, a free market will result in a publicly beneficial outcome. However, a distribution where one person owns everything is Pareto optimal since we cannot make

[15] According to Whyte [2019, p.164], even Friedman thought that Hayek's belief in the market and arguments against any kind of intervention was harmful.

anyone better off without making that person worse off. Thus, we may want to require that the outcome of the market should be more than just Pareto optimal in order to be publicly beneficial. As Jehle and Reny [2011] point out after having compared the first theorem of welfare economics with the invisible hand argument:

> It is extremely important to appreciate the scope of this aspect of competitive market systems. It is equally important to realise its limitations and to resist the temptation to read more into what we have shown than is justified. Nothing we have argued so far should lead us to believe that WEAs [Walrasian equilibrium allocations] are necessarily 'socially optimal' if we include in our notion of social optimality any consideration for matters of 'equity' or 'justice' in distribution. Most would agree that an allocation that is not Pareto efficient is not even a candidate for the socially best, because it would always be possible to redistribute goods and make someone better off and no one worse off. At the same time, few could argue persuasively that every Pareto-efficient distribution has an equal claim to being considered the best or 'most just' from a social point of view [Jehle and Reny, 2011, p.217].

If we want to define publicly beneficial outcomes as outcomes that also satisfy some measures of equality and justice, then a free market will not necessarily lead to it. Further, even if we define Pareto optimal outcomes as publicly beneficial, it is unlikely that any real market will be able to produce such an outcome, since no real market can satisfy the idealised conditions used in the first welfare theorem [Samuels, 2011, p.182]. This observation is especially relevant, since the prisoner's dilemma game (also studied in microeconomics) shows us that under some conditions, choosing to act in accordance with one's own interest - whatever that may be - can lead to a Pareto suboptimal outcome.[16] The result that self-interested actions will lead to Pareto optimal outcomes thus also depends on which microeconomic model we use.

Summing up, we see that the claim that self-interested actions in a market situation can lead to publicly beneficial outcomes has been a part of modern

[16]For a thorough discussion of how prisoner's dilemma disproves the claims made in the invisible hand argument, see Morgan [2012, pp.351-356].

economics since its beginning. Interestingly, its popularity increased from the 1940s both due to its role in neoclassical economics and because of the political tension caused by the second world war and the cold war. Thus, the claim has been a part of both positive and normative economics. Looking at the claim from the perspective of positive economics, we see that it still involves normative judgements. This is both true when considering Hayek's arguments that markets do not exploit people, that they always provide them with opportunities, and that they supply people with what they want. It is also true when we consider how "publicly beneficial outcomes" are defined, since what counts as a beneficial outcome is a matter of normative judgement. This, along with the focus on benefits rather than negative consequences of self-interested actions in a free market suggests that positive economics is not detached from normative judgments as Friedman [1953] claims.

4.4 How commerce can foster individual virtues

Mandeville and Smith both argued that self-interested individual actions can lead to *publicly* beneficial outcomes. Here, I turn to another line of arguments in defence of self-interest and greed. These arguments - made for example by Voltaire [1733] and Hume [1738] - state that self-interest and greed in market situations can lead to *individual* benefits by promoting individual virtues. As before, I first present the historical arguments for the benefits of (narrow) self-interest and then discuss how their legacy can be seen in contemporary microeconomics exemplified by Friedman [1962].

4.4.1 Voltaire's argument for religious tolerance

Voltaire [1733] is one of the first persons to raise the claim that self-interested actions in a market setting can foster virtues among individuals in it. In his *Letters concerning the English nation*, he argues that the difference in religious tolerance between London and Paris is due not to any difference in the dominant churches, but to London's flourishing commerce:

> Take a view of the Royal Exchange in London, a place more venerable than many courts of justice, where the representatives of all nations meet for the benefit of mankind. There the Jew, the Mahometan, and the Christian transact together, as though they all professed the same religion, and give the name infidel to none but bankrupts. There the Presbyterian confides in the Anabaptist, and the Churchman depends on the Quaker's word. At the breaking up of this pacific and free assembly, some withdraw to the synagogue, and others to take a glass. This man goes and is baptized in a great tub, in the name of the Father, Son, and Holy Ghost: that man has his son's foreskin cut off, whilst a set of Hebrew words (quite unintelligible to him) are mumbled over his child. Others retire to their churches, and there wait for the inspiration of heaven with their hats on, and all are satisfied [Voltaire, 1733, letter 6].

According to Voltaire, the London Stock Exchange (and commerce in general) provides a context where people have an interest in putting aside religious differences in order to trade with each other. When the day ends, each merchant can go about their religious duties as they see fit. Thus, commerce promotes religious tolerance due to the economic benefits of dealing with more people. Since the English held their merchants in higher regard than the French, more commerce happened in London, and so London became a place with more religious tolerance [McElroy, 1998, Voltaire, 1733, letter 10].

Promoting religious tolerance

Comparing London and Paris in the 18th century, it is clear that London is more religiously tolerant than Paris, even though the English law and the Church of England did not mirror this tolerance [McElroy, 1998]. Also, Voltaire correctly observes that one main difference between the two cities is the focus on commerce in London. However, Voltaire does not attribute the difference in religious tolerance *solely* to free market powers. First, there are constitutional differences between the two countries. Second, Voltaire observes that the English class structure is less rigid than the French [Voltaire, 1733, letter 9]. Third, England has a comparatively large middle class (due

perhaps to the respect towards trade and the positive economic effects this has for merchants). Finally, England has a greater religious diversity than France. It is with this final point that Voltaire ends his letter considering the religious tolerance one can encounter at the London Stock Exchange:

> If one religion only were allowed in England, the Government would very possibly become arbitrary; if there were but two, the people would cut one another's throats; but as there are such a multitude, they all live happy and in peace [Voltaire, 1733, letter 6].

Thus, while commerce did contribute to an increased religious tolerance in London, Voltaire's other considerations show that a "free market" is not sufficient to explain the difference. Indeed, as I will argue in subsection 4.4.3, markets can also promote intolerance, if this is economically beneficial for the merchants.

4.4.2 Hume's argument for the origin of artificial virtues

Five years after Voltaire's letters, Hume [1738] made a similar argument for the benefits of self-interested actions in a market setting. Contrary to Mandeville, Hume argues that people can be virtuous. We can distinguish between vicious and virtuous actions by considering how they make us feel [Hume, 1738, p.527]. A person or an action is virtuous, if it points to a character trait that - if we consider it from a steady and general point of view - will make us feel sympathy towards it because of the effects such a trait will typically generate throughout a life [Hume, 1738, pp.626,632,636-637].

Next, Hume distinguishes between *natural virtues* and *artificial virtues*. Natural virtues are virtues that we feel a natural sympathy towards because of the motives behind them [Hume, 1738, p.532]. The good resulting from a natural virtue is present in every single act to which it gives rise [Hume, 1738, p.630]. Examples of natural virtues are meekness, beneficence, charity, generosity, clemency, moderation, and equity [Hume, 1738, p.629]. In contrast, artificial virtues of justice - and derived from it; honesty, promise keeping, and respect for property - originate because of a voluntary convention made

among people in a society [Hume, 1738, pp.541-542]. The convention is agreed upon, because everyone realises that it will be beneficial for them to conform to it *provided* that everyone else does the same. Thus, artificial virtues are agreed upon out of self-interest.

In order to explain why artificial virtues - or justice - arise, Hume refers to the three kinds of goods that we want. These are the internal satisfaction of the mind, external advantages of our body, and enjoyment of possessions [Hume, 1738, p.539]. All goods are more easily obtained in a society. However, societies also pose a risk for the enjoyment of possession since property is easily transferable and some resources are scarce. Thus, Hume suggests artificial virtues will emerge, since they allow each person the security to enjoy their property [Hume, 1738, pp.540-541]. He concludes that

> *'tis only from the selfishness and confined generosity of men, along with the scanty provision nature has made for his wants, that justice derives its origin* [Hume, 1738, p.547].

Single acts of justice need not be beneficial for society or for the individual. However, justice in general is beneficial for society and the individuals in it, and it is because of this benefit that people choose to follow the convention [Hume, 1738, pp.548-549,630]. When the convention is established, people will view character traits and actions supporting it with sympathy. Thus, the character traits related to the convention come to be seen as virtues:

> *Thus self-interest is the original motive to the* establishment *of justice: but a* sympathy *with public interest is the source of the* moral approbation, *which attends that virtue* [Hume, 1738, p.551].

On Hume's account then, self-interest together with the scarcity of resources is the origin of artificial virtues in a society. In this way, self-interest can foster individual virtues.

Virtuous individuals

Considering Hume's account of artificial virtues, it is important to realise that they can only be stable if they are backed by law. To see why, consider the public goods game discussed in chapter 3. Everyone will gain the most if

all contribute to the public exchange. However, each individual will benefit the most if they only invest in the individual exchange. Thus, the convention of investing in the public exchange will only be stable if everyone can be certain that all will follow it. One way to ensure this is to enact legislation that punishes people who do not conform to the convention. Thus, a free market is not sufficient for the development and preservation of artificial virtues since - without a rule of law - people will have economic incentives to break them. Further, Hume's argument does not state what kind of market economy is needed. Thus, his argument - like those of Mandeville, Smith, and Voltaire - is compatible with, for example, a system of taxation and redistribution.

4.4.3 Legacy of the "individual virtues" arguments

The argument that self-interested actions in a market setting can foster individual virtues has not received as much attention as the idea that self-interest can lead to public benefits. It is none the less present in contemporary economic thinking and theorizing. To see this, I will focus on Friedman's argument in his famous and widely read book *Capitalism and Freedom* from 1962. The aim of the book is to show that

> increases in economic freedom have gone hand in hand with increases in political and civil freedom and have led to increased prosperity; competitive capitalism and freedom have been inseparable [Friedman, 1962, p.ix].

As a part of his argument, Friedman claims that capitalism reduces discrimination:

> No one who buys bread knows whether the wheat from which it is made was grown by a Communist or a Republican, by a constitutionalist or a Fascist, or, for that matter, by a Negro or a white. This illustrates how an impersonal market separates economic activities from political views and protects men from being discriminated against in their economic activities for reasons that are irrelevant to their productivity - whether these reasons are associated with their views or their color.

> As this example suggests, the groups in our society that have the most at stake in the preservation and strengthening of competitive capitalism are those minority groups which can most easily become the object of the distrust and enmity of the majority - the Negroes, the Jews, the foreign-born, to mention only the most obvious. Yet, paradoxically enough, the enemies of the free market - the Socialists and Communists - have been recruited in disproportionate measure from these groups. Instead of recognizing that the existence of the market has protected them from the attitudes of their fellow countrymen, they mistakenly attribute the residual discrimination to the market [Friedman, 1962, p.21].

Thus, the market is beneficial for people in discriminated groups since they will be able to find employment in activities that are not visible for consumers (such as the production of wheat for bread). Further, a free market will benefit discriminated groups since there will be an economic incentive to separate economic efficiency from other characteristics of the individual [Friedman, 1962, p.109]. To give an example, if it is possible to employ a woman at a lower wage than a man, then it will be economically efficient to employ women rather than men. This also means that employers who do not employ the cheapest workers will induce a cost on themselves such that discrimination will be costly for the ones discriminating [Friedman, 1962, p.110]. Thus, Friedman argues that capitalism reduces discrimination by making it cost effective not to discriminate. Assuming that people's economic motivation is their primary motivation, free markets will therefore eliminate discrimination.

Next, Friedman argues that introducing laws to prevent discrimination will be harmful. He gives three reasons for this position. First, anti-discriminatory laws are a clear interferences "with the freedom of individuals to enter into voluntary contracts with one another" [Friedman, 1962, p.111]. Second, anti-discriminatory laws are equivalent to discriminatory laws (as The Hitler Nuremberg laws) since they both rely on state interventions towards specific groups. Third, the laws will bring harm (solely) to the businesses of people who - without having the bias themselves - merely

respond to the sentiment of a community:

> When the owner of the store hires white clerks in preference to Negroes in the absence of the law, he may not be expressing any preference or prejudice or taste of his own. He may simply be transmitting the tastes of the community. He is, as it were, producing the services for the consumers that the consumers are willing to pay for. Nonetheless, he is harmed, and indeed may be the only one harmed appreciably, by a law which prohibits him from pandering to the tastes of the community for having a white rather than a Negro clerk [Friedman, 1962, pp.111-112].

The shop owner is subjected to *positive* harm by the law since they are coerced into a contract they do not consent to. In contrast, the African American will only experience *negative* harm if the law is not invoked, since they are only harmed by not being able to enter a mutually beneficial contract. This can be compared to an opera singer who lives in a community where people will rather pay for a blues performance. The consumers will also not be harmed since they can find another store with only white clerks.

In summary, Friedman argues, first, that a free market in itself will decrease or even remove discrimination. Second, Friedman argues that anti-discriminatory laws are harmful because they i) intervene with individual freedom, ii) cannot be distinguished from discriminatory laws, and iii) cause positive harm to employers. His first argument can be seen as a modern version of the arguments made by Voltaire and Hume. The second argument can be seen as an argument against the claim that anti-discriminatory laws will reduce discrimination more effectively than a free market.

Is positive microeconomics free from normative judgements? - Free markets and discrimination

Though the aim of *Capitalism and freedom* can be seen as an instance of normative economics, the economic analyses provided by Friedman are cases of positive economics. Thus, it is relevant to ask whether there are any normative judgements in Friedman's positive analyses of how self-interested actions in free markets promote individual virtues by reducing discrimination.

Friedman gives two reasons why free markets reduce discrimination. First, he argues that markets reduce discrimination by providing "unseen" jobs where discriminated groups can find employment. This argument seems to be directed at an alternative scenario where people from discriminated groups will not be able to get a job at all - for example due to a law prohibiting it. While such laws have existed with regard to some jobs - like the requirement that people in the British government in the 18th century had to belong to the Church of England [McElroy, 1998] - it is a far-fetched alternative scenario to consider in 1962. Further, Friedman does not claim that people from discriminated groups will be able to get every job they are qualified for in a free market. Rather, he only states that they will be able to get jobs where they are not seen by consumer. While this might be better than no job at all, the argument does not show how a free market will reduce discrimination among consumers.

Second, Friedman argues that there is an economic incentive for employers in a free market to employ people from discriminated groups both because they demand a lower salary and it increases the number of employees, employers can choose from. It seems reasonable, however, to argue that it is still discrimination if a person is employed at a lower wage solely because of a characteristic that is irrelevant for the performance of the job.[17] Since the shop-owner is assumed to be self-interested, they have no incentive to provide equal wage for their employees if it can be avoided. Further, the increase in supply of workers is only relevant for the shop owner if there is a shortage. As long as there are sufficiently many workers available from the majority group, a discriminating shop owner has no reason to be interested in employees from minority groups.

Since a free market will not necessarily reduce discrimination, it is relevant to ask whether it might foster discrimination by making anti-discriminatory behaviour costly. Friedman [1962, pp.111-112] himself provides an example where it is economically inefficient for a shop owner to be anti-discriminatory in their employments. According to him, a shop owner in a community,

[17] See Altman [2016, section 1.2] and Lippert-Rasmussen [2006] for a definition of discrimination where this is the case.

which harbours a strong bias against African Americans, may experience loses or even bankruptcy if they employ an African American to assist customers. Thus, being non-discriminating can be costly in some contexts. In these contexts, a free market will promote discrimination rather than hinder it, since the shop owner will have to reflect the sentiment of the community in order to stay in business.

Besides his arguments why a free market reduces discrimination, Friedman also gives three reasons why anti-discrimination laws are harmful. First, he states that they are harmful because they are coercive. However, it is unclear how this kind of coercion differs from the coercion imposed by other measures, such as a legal system and minimal government founded by taxation, necessary in order for the free market to work. The only real difference is that the former is motivated by a corrective concern for equality while the latter is not. But unless it can be proved that a free market alone is able correct all our social justice concerns - which I have already argued it cannot - I see no reason why social justice should not be an acceptable aim or motivation for a policy.

Second, Friedman argues that there is no difference between discriminatory laws and anti-discriminatory laws since they are both state interventions aimed at specific groups. Here, it is important to consider the benefits which the laws are intended to produce for said groups. There is a huge difference between laws intended to harm a group and laws intended to help a group. The fact that both types of laws function via state interventions (just like most other laws do) does not make the two types indistinguishably or render it impossible to support the latter without also supporting the former.[18]

Finally, Friedman [1962, pp.112-113] compares the harm done to an opera singer who cannot get employment because people prefer blues to that of an African American who cannot get a job because of their skin colour. This comparison, however, neglects that the opera singer made the choice of singing opera *and* will be able to pursue another career if there is a shortage in demand for opera. In contrast, being born with one skin colour rather than another is not something we choose or can change. Thus, whereas we

[18]See also Lippert-Rasmussen [2006] for a discussion of good and bad discrimination.

might find that there is a grain of individual responsibility in the difficult situation of the opera singer, such individual responsibility does not apply to the African American. This is one reason we may want to correct the disadvantage of the latter without wanting to do the same for former.[19]

Summarising the arguments discussed in this section, we see that the claim that greed and self-interested behaviour in a market setting can foster individual virtues has been around from the beginning of modern economics and can still be found in neoclassical literature. Further, we see that the claim has changed from a general defence of commerce in Hume [1738] and Voltaire [1733] to an argument in defence of a laissez-faire economy with minimal governmental interference. Finally, as I have argued above, Friedman [1962] has a very specific definition of discrimination that does not define - for example - wage differentiation due to colour or gender as discrimination. Even though Friedman's argument is supposed to be based on a purely positive analysis, we thus see how it includes several normative judgements concerning what should count as discrimination or harm. Once again, the line between positive and normative economics is not as clear as Friedman [1953] suggests.

4.5 Conclusion

In this chapter, I have considered Friedman's [1953] claim that we can distinguish between positive and normative economics and that positive economics is independent of normative economics and judgements about how the world ought to be. By looking at the arguments presented by Mandeville, Smith, Voltaire, and Hume, I have shown that economics since its beginning has been strongly influenced by the normative claim that greed and self-interested actions in a market setting is both publicly and individually beneficial. Further, by looking at the 18th century origin of

[19] Using the theory of discrimination proposed by Lippert-Rasmussen [2006, pp.168-169], we also see that African Americans belong to a socially salient group and opera singers do not. If we accept that discrimination can only occur towards socially salient groups, then this again explains the difference between the two cases.

the claims and tracing their legacy to neoclassical economics, we see how the descriptive assumption of how humans act is closely connected to the normative claim that it is beneficial to act in accordance with one's (narrow) self-interest. Indeed, it is easy to see how an if-then clause like "if we want to control people's passions, then promoting greed is the best way to do it" can become a normative clause by simply adding the premise "and we want to control people's passions". Finally, we see how the neoclassical arguments - presented as positive economics - still use very specific and value-laden definitions of, for example, "publicly beneficial" and "discrimination" that we might not want to accept. Thus, positive economics is not independent of normative judgements, as Friedman [1953] claims.

Evaluating the arguments presented in this chapter - and as the empirical findings discussed in chapter 3 indicate - promoting greed and self-interested behaviour will not necessarily produce a beneficial outcome for society or individuals. Since this is already exemplified within economic theory (for example by the prisoner's dilemma game), one can ask why the idea that humans act, and ought to act, in accordance with their (narrow) self-interest is still present in microeconomics textbooks. I will consider this question in the next chapter, where I first describe the rationality assumption's historical development and then present its current variants in the standard models as they are described in microeconomics textbooks.

Chapter 5

Rationality in microeconomics

5.1 Introduction

In chapter 4, we saw how the claim that (narrowly) self-interested actions are beneficial has been a part of modern economics since its beginning and is still part of the positive arguments provided by neoclassical economists and presented in microeconomics textbooks today. Evaluating the arguments for the benefits of self-interested actions, however, we also see that such actions are not necessarily beneficial. This raises the question of why the assumption that people act to maximise their own (monetary) gains is present in microeconomics textbooks and teaching today.

The aim of this chapter is to show why and how neoclassical economics assumes that agents act rationally. The chapter is divided into three parts. First, I provide a historical account of the rationality assumption and argue that its development is closely related to the development of economic theory. Second, I present the current variants of the assumption used in the standard models in microeconomics textbooks along with the informal discussions accompanying them. These discussions clearly suggest that we act, or ought to act, rationally. Thus, even though economics textbooks generally support the methodological instrumentalism of Friedman [1953] -

stating that assumptions such as the rationality assumption are false and do not describe how people act or ought to act - they violate its requirements by claiming that people both act and ought to act rationally. We also see that even when the rationality assumption can be satisfied by agents with other-regarding preferences, the textbooks assume that agents are (narrowly) self-interested. In the final part of the chapter, I use the above findings to discuss why the rationality assumption is such a prominent assumption in microeconomics textbooks today.

The chapter thus provides the background analysis needed to state the main argument that microeconomics can be self-fulfilling by promoting a social norm of (narrow) self-interest via the dissemination of textbooks and university teaching practices.

5.2 How the rationality assumption came to be

In this section I argue that the development of the rationality assumption played a crucial role in the development of contemporary microeconomics. I do this by showing how the assumption was consciously created through a series of idealisations and exaggerations intended to separate economics from other scientific disciplines and to enable the change from a verbal to a formal science.

The section is divided into four subsections that chronologically trace the development of the assumption. In each subsection I present the arguments made by the relevant economists, and discuss how their accounts of human behaviour relate to the general development of economic theory.[1]

5.2.1 Smith: a non-systematic account of human behaviour

In order to appreciate the development of the rationality assumption, I start with Adam Smith's account of human behaviour in *Wealth of nations*.[2]

[1] The following account of how the rationality assumption developed is based on Morgan [2012, ch.4] and further supported by my analyses of secondary and primary literature.

[2] I will assume that his account of human behaviour in *Moral sentiments* is consistent with that in *Wealth of nations*. The idea that the two accounts are contradictory - known as

Rather than providing one stringent account, Smith unsystematically refers to several different human propensities, preferences, motivations, virtues, vices, passions, and habits in order to make different arguments for how humans interact and the economy works.

Smith's first characterisation of human behaviour is found when he discusses three benefits of division of labour. First, people who only do one thing will improve their skills with regard to this thing. Second, it will save time since people do not have to change tasks:

> A man commonly saunters a little in turning his hand from one sort of employment to another. When he first begins the new work he is seldom very keen and hearty [...] The habit [...] renders him almost slothful and lazy, and incapable of any vigorous application even on the most pressing occasions [Smith, 1776, p.16].

Finally, people who only consider one task have more mental capacity available to consider how to improve it. This narrow focus, combined with a desire to enjoy more spare time, will make people more likely to invent new machinery and improve the production processes [Smith, 1776, p.17].

Next, Smith considers how division of labour is possible. Here, he notes that unlike animals, humans have the faculties of speech and reason which induce a propensity to truck, barter, and exchange [Smith, 1776, p.21]. This propensity enables people to form contracts and cooperate. However, since people depend so much on one another, it is beneficial to make it in the self-interest of others to interact with us:

> But man has almost constant occasion for the help of his brethren, and it is in vain for him to expect it from their benevolence only. He will be more likely to prevail if he can interest their self-love in his favour, and shew them that it is for their own advantage to do for him what he requires of them. Whoever offers another a bargain of any kind, proposes to do this. Give me that which

the *Adam Smith problem* - is likely to have originated from the misconception that *Wealth of nations* depicts humans as *only* self-interested while *Moral sentiments* depicts them as benevolent [Fitzgibbons and Fitzgibbons, 1995, pp.3-4]. For papers supporting my position, see for example Macfie [1959], Herbener [1987], Fitzgibbons and Fitzgibbons [1995], Sen [1997], Montes [2008], Paganelli [2009], Moene [2011], Campbell [2014], and Brady [2018].

> I want, and you shall have this which you want, is the meaning of every such offer; and it is in this manner that we obtain from one another the far greater part of those good offices which we stand in need of. It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love, and never talk to them of our own necessities but of their advantages. Nobody but a beggar chuses to depend chiefly upon the benevolence of his fellow-citizens [Smith, 1776, p.22].

Thus, in order to prevail, it is beneficial not only to ask for help, but also appeal to a mutual advantage that two people can get by trading. Smith ends his discussion on division of labour by telling a story of how people come to have different occupations: in a tribe of hunters, people will end up performing the tasks they have the most talent for, since it will increase their opportunities for exchanging their services with others. However, the difference in talents between humans is actually very small, and so differences between people arise not so much because of nature as because of habit, custom, and education [Smith, 1776, pp.23-24]. Thus, Smith's arguments concerning division of labour already provide a complex description of human characteristics and behaviour which is not limited to a self-interested pursuit of monetary gains.

Smith also discusses human behaviour when he addresses how people accumulate and employ their capital. First, people try to work in ways that increase their capital. Once they have sufficient funds to maintain themselves for months or years, they will seek to use their additional capital to create revenue [Smith, 1776, pp.160,162]:

> In all countries where there is tolerable security, every man of common understanding will endeavour to employ whatever stock he can command in procuring either present enjoyment or future profit. [...] A man must be perfectly crazy who, where there is tolerable security, does not employ all the stock which he commands, whether it be his own or borrowed of other people, in some one or other of those [...] ways [Smith, 1776, p.169].

The passion for present enjoyment is violent and alternating, and people will want to invest in it:

> Thus, not only the great landlord or the rich merchant, but even the common workman, if his wages are considerable, may maintain a menial servant; or he may sometimes go to a play or puppet-show [Smith, 1776, p.194].

However, everyone is also born with an ever-present desire to better their conditions and this will prompt them to save their money [Smith, 1776, pp.203,205]. Through a lifetime, the desire to better one's conditions will be greater than the passion for present enjoyment, and so people will save rather than spend [Smith, 1776, p.204]. Smith also recommends investing in durable things that will maintain their value - such as buildings, furniture, books, statues, pictures, or jewels - since this will make a person richer than investing in things that do not last - such as servants, food, dogs, and horses [Smith, 1776, pp.209,210]. Investing in durable things will further enable a person to adjust their spending without humiliation since it will not be noticed by the public or their neighbours [Smith, 1776, p.211]. Finally, Smith discusses whether to spend money on a hospital - which will not produce any goods and thus will not create value (due to the labour theory of value) - or whether to spend money on commodities such as jewels and pictures (which will produce goods and thus create value). Here he writes:

> The latter species of expence, therefore, especially when directed towards frivolous objects, the little ornaments of dress and furniture, jewels, trinkets, gewgaws, frequently indicates, not only a trifling, but a base and selfish disposition [Smith, 1776, p.212].

Thus, we again see that Smith has a very broad account of human beings, and that he in some cases praises benevolent actions and scorns selfish ones.

Last, Smith describes human behaviour when he analyses the progress of wealth in different nations. Here, he argues that humans have a natural inclination to cultivate land and live in the countryside:

> The beauty of the country besides, the pleasure of a country life, the tranquillity of mind which it promises, and wherever the

> injustice of human laws does not disturb it, the independency which it really affords, have charms that more or less attract every body; and as to cultivate the ground was the original destination of man, so in every stage of his existence he seems to retain a predilection for this primitive employment [Smith, 1776, p.229].

Cultivating land is attractive because it makes people feel they are independent and masters of their own lives [Smith, 1776, p.231]. As already mentioned in section 4.3, Smith uses this desire to argue that city merchants will buy up land in the countryside [Smith, 1776, p.259]. Finally, the noblemen who once had all the power and wealth, will use it "for the gratification of the most childish, the meanest and the most sordid of all vanities" and thus "gradually [barter] their whole power and authority" [Smith, 1776, p.264].

In summary, Smith presents a complex and fragmented picture of human beings, which cannot easily be used to predict human behaviour or make a model of human actions. Since economics was a verbal science at the time, he uses analogies, stories, and historical episodes to argue how human behaviour influences the economy in different ways.[3] Though Smith states that people can rely on market exchanges because they are mutually beneficial, he does not believe that people *only* act out of self-interest or even that it would be desirable if people did so.

5.2.2 Malthus and Mill: idealised economic behaviour

22 years after the publication of *Wealth of nations*, Malthus [1798] provided a simple argument for why population growth will be stopped by people's need for food:

> I think I may fairly make two postulata.
> First, That food is necessary to the existence of man.
> Secondly, That the passion between the sexes is necessary and will remain nearly in its present state. [...]
> Population, when unchecked, increases in a geometrical ratio. Subsistence increases only in an arithmetical ratio. A slight

[3] It is interesting that Smith in his book often do not consider people in general, but rather refers to characteristics of people in specific occupations or social classes [Smith, 1776, pp.209-210, 229-230, 259-260].

> acquaintance with numbers will shew the immensity of the first power in comparison of the second. [...]
>
> This implies a strong and constantly operating check on population from the difficulty of subsistence. This difficulty must fall somewhere and must necessarily be severely felt by a large portion of mankind [Malthus, 1798, pp.4-5].

By looking at birth rates and child mortality, Malthus [1798, p.7] argues that the population will be doubled every 25 years. In contrast, he argues that the production of food in the world cannot increase at the same rate. Thus, the amount of food available will set a natural (and very unpleasant) boundary for population growth unless people are able to let their reason control their passion for reproduction.

Malthus's argument is based on a simple model for population growth and food production. His model can be seen as an early example of an *Aristotelian idealisation* where features of human motivations are removed in order to get a simple account of human behaviour [Frigg and Hartmann, 2018]. By focusing on two motives only - i) the instinct to reproduce and ii) the reason not to bring children to the world that cannot be provided for - Malthus is able to explore what will happen if the instinct is subdued to reason and vice versa.

Malthus's model of population growth worked as an exemplar for Mill's argument in 1836, concerning the definition of economics and the scope of its enquiry [Morgan, 2012, p.139]. Here, Mill defines economics as:

> The science which traces the laws of such of the phenomena of society as arise from the combined operations of mankind for the production of wealth, in so far as those phenomena are not modified by the pursuit of any other object [Mill, 1836, p.99].

This definition distinguishes economics from several other branches of human knowledge. First, economics is a science rather than an art because it is concerned with a collection of *truths* aimed at discovering *laws* rather than a body of particular *rules* [Mill, 1836, pp.88-89]. Second, economics is a moral science rather than a physical science. Physical sciences are concerned with the *laws of matter*: how different objects react and interact. Moral sciences

are instead concerned with the *laws of mind*: human intentions and actions [Mill, 1836, pp.92-92]. Economics is a moral science "which treats of the production and distribution of wealth, so far as they depend upon the laws of human nature" [Mill, 1836, p.94]. Finally, Mill distinguishes economics from other moral sciences by considering three categories of laws of mind. These are, first, laws concerning people with no social interactions. Second, laws concerning feelings and affections we have in relation to other people, and, third, laws concerned with the feelings and ideas of people living together in a society for common purposes [Mill, 1836, pp.95-96]. According to Mill, economics is only concerned with laws of the third kind. However, there are several moral sciences concerned with man living in society. Mill therefore narrows the areas of human nature considered in economics even further:

> [Economics] is concerned with him [humans] solely as a being who desires to possess wealth, and who is capable of judging of the comparative efficacy of means for obtaining that end. It predicts only such of the phenomena of the social state as take place in consequence of the pursuit of wealth. It makes entire abstraction of every other human passion or motive; except those which may be regarded as perpetually antagonizing principles to the desire of wealth, namely, aversion to labour, and desire of the present enjoyment of costly indulgences. [...] Under the influence of this desire, it shows mankind accumulating wealth, and employing that wealth in the production of other wealth [Mill, 1836, pp.97-98].

Thus, in order to distinguish economics from other moral sciences, Mill removes all but three human desires and defines economics as considering how the production of wealth occurs in a society *assuming* that people only desire to:

- possess wealth (and always prefer more to less),
- avoid labour, and
- spend wealth on present enjoyments.

Mill is aware that by removing most human desires, economics will not consider how real humans behave:

> Not that any political economist was ever so absurd as to suppose that mankind are really thus constituted [...] There is, perhaps, no action of man's life in which he is neither under the immediate nor under the remote influence of any impulse but the mere desire of wealth [Mill, 1836, p.98].

Why, then does Mill propose to use such a definition of humans in economics? According to Mill, if we want to understand a phenomenon which occurs due to several different causes, we first have to study and understand the isolated effects of each of the causes. Once we know the effect of each cause in isolation, we can then turn to the study of the effects of multiple causes simultaneously. Economics, then, is an abstract science that assumes a simplified definition of human nature in order to study one cause (the desire for wealth) in isolation [Mill, 1836, p.101]. Since this assumption is not necessarily founded in facts, economists cannot say anything about what will actually occur in the real world (where people are affected by several causes). Instead, economics focuses on what *tendencies* one cause might produce, where this cause can be more or less present in real world situations [Mill, 1836, pp.102-103,111,113].

In summary, Mill defines economics as a distinct science from other moral sciences by reducing its scope to phenomena related to the production of wealth and the three human desires related to it. Thus, Mill makes an Aristotelian idealization of human nature *in order to* define economics as a separate science. The simplification of human nature enabled both Malthus and Mill to reason about the effects of a few human desires. However, it comes at the price that economics can no longer be applied directly to the world.[4]

5.2.3 Menger and Jevons: idealisation and exaggeration

The theory of human behaviour in economics was further advanced during the marginal revolution in the 1870s. Here, Jevons [1871], Menger [1871], and

[4]The fact that economics, according to Mill, can only produce tendency laws does not mean that it is obsolete. On the contrary, these tendency laws are relevant for explaining everyone's behaviour, as long as other causes are allowed for as well [Morgan, 2012, p.141].

Walras [1874] independently proposed a marginalist theory of value focused on consumption rather than production [Moscati, 2018, p.25]. As a part of this shift, both Menger and Jevons contributed to the development of the rationality assumption [Morgan, 2012, pp.141-168].

Menger starts his economic theory by *imagining* a situation where people only engage in economic activities [Morgan, 2012, p.142]. In this situation, Menger argues that people will strive to satisfy their needs by acquiring goods [Menger, 1871, p.77]. People's needs are different, and whether they can be satisfied will depend on their individual circumstances [Menger, 1871, p.90,114]. Since the availability of goods can vary, different goods can obtain different significance for people. Menger calls this significance value:

> Value is thus the importance that individual goods or quantities of goods attain for us because we are conscious of being dependent on command of them for the satisfaction of our needs [Menger, 1871, p.115].

Some goods - such as food - are important for maintaining life, while other goods are important for satisfying different degrees of well-being, or even passing enjoyments [Menger, 1871, p.123]. Thus, goods can satisfy needs of different kinds. Further, one need can be satisfied to a larger or smaller degree [Menger, 1871, p.124]. To see this, consider our need for food. We need food in order to survive. However, if we have more food than we need, additional food will not be important for us. Moreover, we will reach a point where additional quantities of a good (like food) will become a burden and cause pain [Menger, 1871, p.125]. Thus, as the importance of one good diminishes, another good will become more important to the individual, and so they will try to acquire this good instead. Finally, Menger stresses that human knowledge is not perfect, and so it is possible for people to err with regard to the importance (and value) of a good [Menger, 1871, p.120]. In the same line of reasoning, Menger argues that people seldom try to obtain exact knowledge of their possessions. Instead, they are satisfied with a degree of exactness that is needed for their practical purposes [Menger, 1871, p.90].

In summary, Menger - like Mill - uses an Aristotelian idealisation to account for human behaviour in his economic theory. However, unlike Mill

(who starts with real people and then discards all but three desires), Menger simply imagines a world where people *only* engage in economic activities. In this imagined world, he argues that human behaviour is constituted by each person's desire to satisfy their needs. Thus, where Mill only looks at a part of reality, Menger removes his account of human behaviour from reality.[5] Finally, Menger's economic theory examines how people will *choose* the goods that - given their circumstances - best satisfy their needs. As we shall see in the next subsection, this emphasis on choices rather than internal motivations is an important step in the development of the rationality assumption.

Whereas Menger provides a verbal account of human behaviour, Jevons sees economics as dealing with quantities, and therefore as purely mathematical in character [Jevons, 1871, p.3].

> The theory which follows 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].

Jevons defines pleasure and pain with reference to Bentham's seven circumstances that influence the pleasure we get from an act or feeling. These are intensity, duration, uncertainty, remoteness, fecundity, purity, and extent (how many people are affected by it) [Jevons, 1871, pp.28-29]. Jevons acknowledges the importance of all seven circumstances, but he restricts the economic account of pleasure to *intensity* and *duration* so that he can represent it by a graph [Jevons, 1871, p.29]. Letting the y-axis displays the intensity of a feeling as a function of time on the x-axis, he argues that the graph will be continuous and decrease over time. Pain is represented as negative intensity since it is the opposite of pleasure [Jevons, 1871, p.32]. Finally, we can account for the uncertainty of future events by estimating the probability of different events occurring and weight the pleasure of an event by its probability of occurring [Jevons, 1871, pp.35-36]. Here, we get another glimpse if Jevons's account of human behaviour:

[5] For a further comparison and discussion of Mill's and Menger's theories, see Cartwright [1994] in Hamminga and De Marchi [1994, pp.171-188], Zouboulakis [2001], and Morgan [2012, pp.141-145]. See Menger [1883] for Menger's own discussion of his method.

> Almost unconsciously we make calculations of this kind more or less accurately in all the ordinary affairs of life; and in systems of life, fire, marine, or other insurance, we carry out the calculations to great perfection. In all industry directed to future purposes, we must take similar account of our want of knowledge of what is to be [Jevons, 1871, p.36].

Having introduced the basic character of pleasure and the human desire to maximise it, Jevons turns to the topics studied in economics. First, he defines *commodity* as anything that can give pleasure or prevent pain and *utility* as the quality that makes an object a commodity by giving us pleasure or pain [Jevons, 1871, pp.37-38]. Thus, utility is a relation between goods and humans' requirements that is measured by the increase in happiness a person experiences from a good [Jevons, 1871, p.43]. How much utility a person gets from an additional amount of a good will depend on how much of the good the person already has [Jevons, 1871, pp.45-46]. Notably, Jevons, like Menger, argues that when the quantity of a good increases, its utility for us can drop to zero [Jevons, 1871, p.53]. Second, Jevons suggests that utility can be represented by a graph with the quantity of a good on the x-axis and the utility of the consumer on the y-axis [Jevons, 1871, p.47]. This enables him to produce the first *continuous utility curve* by arguing that we *theoretically* can imagine quantities of goods being infinitely divisible [Jevons, 1871, pp.47-48]. Finally, Jevons turns to the question of exchange. Here, he defines the value or purchasing power of a good as the ratio of exchange for other commodities [Jevons, 1871, pp.79-84]. Further, he defines the market studied in economics as a *perfect market*.

> By a market I shall mean two or more persons dealing in two or more commodities, whose stocks of those commodities and intentions of exchanging are known to all. It is also essential that the ratio of exchange between any two persons should be known to all the others. It is only so far as this community of knowledge extends that the market extends. [...] Every individual must be considered as exchanging from a pure regard to his own requirements or private interests, and there must be perfectly free competition, so that any one will exchange with any one else for the slightest apparent advantage. There must be no conspiracies

> for absorbing and holding supplies to produce unnatural ratios of exchange [Jevons, 1871, pp.85-86].

Since everyone in the market has complete knowledge of supply and demand, each commodity will only have one price. Given this, Jevons introduces his *theory of exchange* in which a person will trade as long as they can increase their utility, and stop trading when all available trades will lead to a loss of utility [Jevons, 1871, p.96]. Thus, Jevons concludes:

> The general result of exchange is thus to produce a certain equality of utility between different commodities, as regards the same individual; but between different individuals no such equality will tend to be produced. In Economics we regard only commercial transactions, and no equalisation of wealth from charitable motives is considered. [...] But so far as is consistent with the inequality of wealth in every community, all commodities are distributed by exchange so as to produce the maximum of benefit. [...] No one is ever required to give what he more desires for what he less desires, so that perfect freedom of exchange must be to the advantage of all [Jevons, 1871, p.141].

In summary, Jevons - like Mill and Menger - uses Aristotelian idealisations to reduce the complexity of human nature considered in economics. He does this in order to represent economics mathematically. First, he assumes that people *only* strive to *maximise* their own pleasure or utility since it is mathematically easy to find optima. Second, he reduces the definitions of pleasure and utility to two dimensions so that they can be depicted by graphs. Further - and contrary to Menger - Jevons also *exaggerates* some human features. First, he assumes that humans always make their decisions (at least unconsciously) by *calculating* what will give them the highest utility. Second, Jevons assumes that people in a perfect market are able to collect, comprehend, and store an infinite amount of information and knowledge. Thus, Jevons not only uses Aristotelian idealisations, but also *Galilean idealisations*, where features of human nature are distorted - and here exaggerated.[6] Finally, we see how Jevons - like Menger - moves away from considering (a part of)

[6]For a similar discussion, see Morgan [2012, pp.145-150].

the real world and instead looks at a market that does not exist in reality, inhabited by individuals with a single-minded purpose and exaggerated abilities.

Jevons's argument that economics is purely mathematical in character was much debated at the time. However, by moving from a verbal representation of economics to a formal one, Jevons introduced a powerful reasoning tool that came to have a huge impact on the development of economics as we know it today.

5.2.4 Knight: Creating the rationality assumption

In 1929, Knight created the version of human behaviour now used in neoclassical economics [Morgan, 2012, p.150]. As with Jevons and Menger, Knight starts his analysis of perfect competition by considering human behaviour:

> It [economics] assumes that men's acts are [...] directed toward the "satisfaction of wants" [Knight, 1921, p.52].

Knight defines *utility* as the power of things to satisfy wants [Knight, 1921, p.61]. There is no absolute measure or standard of utility since it only makes sense in relation to alternative choices [Knight, 1921, p.63]. Turning to his theory of exchange, Knight argues that "it will be necessary to simplify the situation as far as possible by a process of "heroic" abstraction" [Knight, 1921, p.76]. The idealisations and abstractions relevant for us are summarised below [Knight, 1921, pp.76-81]:

- People are rational and know all consequences of potential acts so that they seek what they want in the light of this knowledge.
- Each person is free to define and follow their own interests. There is no coercion from society or other people in it.
- Every individual is independent of other people; free from social wants, prejudices, preferences, repulsions or any other value not completely manifested in trading.
- There is no cost associated with moving or changing plans.

- Labour and commodities are assumed to be infinitely divisible.
- The production of goods is constant and continuous with no breakdowns. Each person produces one good and the good is consumed instantly.
- The means for productions are all a part people's endowment along with the materials they each need for production. It is assumed that the skills of each individual is fixed and cannot change.
- The only way to acquire a good is through trading. Exchange of goods is instantaneous and costless. There is no fraud, deceit, or theft.
- People are motivated to engage in exchange.
- There is perfect competition with perfect, continuous, costless communication between everyone. Everyone knows instantly what to choose among offers and from whom.
- There are no collusions or monopolies.
- There will be no fluctuation or progression in the market. No factors and conditions will change unless this is explicitly stated and considered.

Here, we see that Knight both reduces the economic account of human motivations and exaggerates their ability to collect and use information and to calculate how best to trade. Thus, Knight - as Jevons - uses both Aristotelian and Galilean idealizations. The result is a caricature of human behaviour that can never occur in reality [Gibbard and Varian, 1978, Morgan, 2012, pp.158-159].

Interestingly, Knight is conscious about the distorted picture of humans which he presents:

> The above list of assumptions and artificial abstractions is indeed rather a formidable array. The intention has been to make the list no longer than really necessary or useful, but in no way minimize its degree of artificiality, the amount of divergence of the hypothetical conditions from those of actual economic life about us [Knight, 1921, p.81].

Knight argues none the less that it is useful to study such idealised situations. By looking at a pure economic situation, we can learn about laws also present in real economic situations. If these laws are sufficiently dominant compared to other factors in the real situation, the analysis of the idealised situation can provide us with a picture of what the real situation may tend towards [Knight, 1921, pp.1-5]. This said, Knight stresses that the results obtained in economics *cannot* be applied to the real world.[7] Indeed, he argues that economists have failed to make it clear that the results they get under idealized conditions cannot be directly applied to reality:

> The limitations of the results have not always been clear, and theorists themselves as well as writers in practical economics and statecraft have carelessly used them without regard for the corrections necessary to make them fit concrete facts. Policies must fail, and fail disastrously, which are based on perpetual motion reasoning *without the recognition that it is such* [Knight, 1921, p.11].

It is because of this that Knight so explicitly states all the assumptions he makes about human behaviour [Knight, 1921, p.11].

Knight's account of human behaviour - combined with his definition of complete markets and equilibrium theory - makes it possible to use mathematics to explore what will happen in this idealised situation. It also shows us how the different accounts of human behaviour developed during the marginal revolution are merged into one theory. As Jevons, Knight uses mathematics and assumes that people have exaggerated knowledge and strive to maximise their utility by calculating what the best outcome is. However, Knight - as Menger - does not focus on the pleasure people derive from a good. Instead, he explains that utility only makes sense as a relative concept when we consider how to *choose* between alternative goods. Further, Knight uses *rationality* to describe "consistent choosing" rather than "reasoning about means to an end" (as Hume [1738] used the term). This supports the claim that the term has gradually changed its meaning in

[7] Knight's methodology may be compared to Weber's methodology of ideal types, also developed in the first half of the 20th century [Kim, 2019].

economics from the latter to the former [Morgan, 2012, p.155]. As we shall see in the next section, one reason the focus on choices rather than introspection is important in microeconomics is that it enables us to *observe* rational behaviour whenever people make logically consistent choices [Morgan, 2012, p.156]. It has also enabled neoclassical economists not only to say that this is how an idealised agent would choose, but to argue that the theoretical results can be applied to real situations where people choose consistently, and even to make the normative claim that people ought to choose in this way.

Next, I turn to the current account of human behaviour in microeconomics textbooks as it is presented in the different variants of the rationality assumption in the standard microeconomic models. I will return to the relationship between economic theory and the rationality assumption in the end of the chapter.

5.3 Current variants of the rationality assumption in microeconomics textbooks

Today, the rationality assumption is an integrated part of almost all microeconomic models taught at universities. Since the models have different subject matters, the formulation of the rationality assumption can vary from model to model. Common for all variants, however, is that they include the two claims

1 that rational agents *only* care about their individual gains (whatever they might be) and

2 that rational agents strive to *optimize* their *choices* with regard to their individual gains.

The variants differ in their definition of *individual gains* which need not be monetary and - in some cases - can allow for other-regarding preferences. Here, I survey three microeconomics textbooks used in the curriculum at universities around the world, with the aim of providing an accurate picture of the arguments and justifications written in microeconomics textbooks with

regard to the rationality assumption for standard microeconomic models. The textbooks are Varian [2014], Jehle and Reny [2011] and Mas-Colell et al. [1995].[8] By using texts excerpts from the three textbooks as examples, I present the different variants of the rationality assumption - along with the informal discussions accompanying them - as they occur in

- consumer theory and theory of the firm,
- choice under uncertainty, game theory, and social choice theory, and
- market behaviour and general equilibrium.

The topics are chosen because they are part of the standard microeconomics syllabus and thus covered by all three textbooks. The detailed analysis of the rationality assumption, conveyed in this section, will be a reference point for the remaining chapters in the thesis, since it provides the background needed for making a precise argument for how microeconomic theory can change people's behaviour by promoting a social norm of (narrow) self-interest.

5.3.1 Consumer theory and theory of the firm

Consumer theory

The obvious place to start the analysis of human behaviour in microeconomics textbooks is with the theory of consumption. This theory states how a consumer will act in a market setting when choosing between different goods. According to Jehle and Reny [2011, p.3],

[8] Varian [2014] is used as an introduction to microeconomics at most UK universities [Earle et al., 2016, p.45] including, for example, LSE (EC201 and EC440) and Warwick university (EC204). Varian [2014] is also used as an introductory textbook to microeconomics at Copenhagen university. Jehle and Reny [2011] is for example used as the primary textbook at Warwick (EC9A1), UCSB (210A), UCL (GLBH0030) and UNIL (microeconomics). Mas-Colell et al. [1995] corresponds to Jehle and Reny [2011] but is older, more well-known, and more mathematical. It is used as textbook at LSE (EC487 and EC411), WSU (EconS501), and Copenhagen University (microeconomics 1). Further Mas-Colell et al. [1995] and Jehle and Reny [2011] are recommended as the best textbooks for microeconomics at http://econphd.econwiki.com/books.htm. Together with Varian [2014] they span microeconomics teaching at universities for both undergraduate and graduate level.

> [t]here are four building blocks in any model of consumer choice. They are the consumption set, the feasible set, the preference relation, and the behavioural assumption.

The consumption set is the set of possible consumption bundles (bundles of goods) that the consumer can choose from. The feasible set is the set of consumption bundles that the consumer can afford. The preference relation is a binary relation that allows for comparison of two goods, by saying one good, x, is at least as good as the other good, y. Finally, the behavioural assumption is that "the consumer seeks to identify and select an available alternative that is most preferred in the light of his personal tastes" [Jehle and Reny, 2011, p.4].

Here, it is important to note the narrow scope of consumer theory. Unlike a general account of decision theory [Savage, 1972, Jeffrey, 1990], consumer theory is *only* concerned with how one consumer chooses between different bundles of goods. Since there is only one individual in the model, and since the question is - simply put - whether this individual prefers to *consume* oranges or apples, it does not make sense to talk about other-regarding preferences within the scope of the model.

In order to use the model, a couple of mathematical constraints are introduced. First, the consumption set, X, is the set of all non-negative bundles of goods $X \subseteq \mathbb{R}^n_+$. X is closed, convex, and $0 \in X$. Second, the consumer's preference relation is assumed to be *rational* by being *complete* (so that for all comparisons of two consumption bundles in the consumption set, one will be preferred to the other), and *transitive* (if x is preferred to y and y is preferred to z, then x should be preferred to z) [Mas-Colell et al., 1995, p.6].[9]

When describing completeness, Mas-Colell et al. [1995, p.6] states that by introspection we know that it can be hard to evaluate alternatives and that it takes time to do so:

> The completeness axiom says that this task has taken place: our decision makers make only meditated choices [Mas-Colell et al., 1995, p.6].

[9] Varian [2014, p.35] also mentions reflexivity (that a good, x is at least as good as itself).

Jehle and Reny [2011, p.5] do not comment on the completeness axiom, and Varian [2014, p.35] writes that the completeness axiom "is hardly objectionable".

Regarding transitivity, Mas-Colell et al. [1995, pp.6-7] argue that it is a strong assumption (as evident by introspection) but that it is "fundamental in the sense that substantial portions of economic theory would not survive if economic agents could not be assumed to have transitive preferences". Jehle and Reny [2011, pp.5-6] argue that this is the assumption which ensures that consumers' choices are *consistent*. At first point, it seems *simple* and *natural* to assume, however it is *controversial* since experiments have shown that peoples choices are not always transitive. Finally, Varian [2014, pp.35-36] writes that the transitivity assumption is not logically *necessary*, but that it is an assumption about people's choice behaviour:

> What would you think about a person who said that he preferred a bundle X to Y, and preferred Y to Z, but then also said that he preferred Z to X? This would certainly be taken as evidence of peculiar behavior.
>
> More importantly, how would this consumer behave if faced with choices among the three bundles X, Y, and Z? If we asked him to choose his most preferred bundle, he would have quite a problem, for whatever bundle he chose, there would always be one that was preferred to it. If we are to have a theory where people are making "best" choices, preferences must satisfy the transitivity axiom or something very much like it. If preferences were not transitive there could well be a set of bundles for which there is no best choice [Varian, 2014, p.36].

Here then, Varian [2014] uses a normative argument to defend a descriptive assumption: because it would be beneficial for individuals to have transitive preferences, we can assume that people do have transitive preferences.

Besides the conditions for rational preferences, it is assumed that people have *well-behaved* preferences. Well-behaved preference relations satisfy the following:

- *Continuity* such that there are no "jumps" in the preferences.

- *Strict monotonicity* (or less demanding: *local non-satiation*) so that wants are unlimited, or more is always better than less.
- *Strict Convexity* (or less demanding: *convexity*) so that the preference for an additional amount of one good decreases as a function of the quantity the person already has. This ensures that people will always prefer to have a combination of goods rather than only having one good.

The preference relation introduced above is often represented by a utility function. A utility function, $u : X \to \mathbb{R}$, represents a rational preference relation, $\succsim$, if for all $x, y \in X$, $x \succsim y \Leftrightarrow u(x) \geq u(y)$ [Mas-Colell et al., 1995, p.9]. Further, if a preference relation is well-behaved there exists a utility function that represents it [Jehle and Reny, 2011, p.14]. Notice that the definition of a utility function means that the functions only carry ordinal value so that the actual numbers in the function convey no meaning about the intensity of a preference.

Finally, it is assumed that people *act rationally* by always choosing the most preferred alternative available to them. This assumption is stated, but not discussed in Jehle and Reny [2011] and Mas-Colell et al. [1995]. Varian [2014, p.3] provides the following comment:

> [This] is *almost* tautological. If people are free to choose their actions, it is reasonable to assume that they try to choose things they want rather than things they don't want. Of course there are exceptions to this general principle, but they typically lie outside the domain of economic behavior.

Based on this assumption, it is possible to formulate the *consumer's problem* or the *utility-maximisation problem*. Assume that we have a market economy with fixed prices, p_i, for every good $x_i \in X$ and each consumer is endowed with a fixed money income $y \geq 0$. The consumer's problem is to maximise their utility given their budget:

$$\max u(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{p} \cdot \mathbf{x} \leq y \tag{5.1}$$

Where **x** and **p** are vectors of goods and corresponding prices. If all assumptions above are satisfied, there is a unique maximum for the problem [Jehle and Reny, 2011, p.21]. Further, if the utility function is assumed to be differentiable, the problem can be solved using the Lagrange method from mathematics.

This introduction of consumer theory is based on a theory about people's preferences. Unfortunately, we only have access to our preferences through introspection [Mas-Colell et al., 1995, p.5]. In order to avoid this limitation, it is standard for microeconomics textbooks to end the chapter(s) on consumer theory with a discussion of *revealed preference theory* [Samuelson, 1948b]. This theory shows that consumer *choices* can be used as a basis for the consumer's problem if they satisfy a series of axioms which ensure that the revealed preferences can be represented by a well-behaved preference relation as discussed above.[10]

In summary, consumer theory analyses people who have well-behaved preferences, and choose between bundles of goods such that they maximise their own utility given their income. Thus, they satisfy the rationality assumption by having rational preferences and acting rationally. Since consumer theory does not account for relations to other people or preferences besides what one wishes to consume, people in the model are completely self-interested.[11] Further - and in contrast to Menger's and Jevons's theories - consumers' wants are unlimited. Notice finally that since consumers are assumed to maximise their utilities by buying goods, and since it is assumed that more goods will always increase their utility, people will always want to increase their income, since this will increase their budget and thus automatically their utility. Thus, the model also assumes that consumers are *narrowly* self-interested.

[10] Different axioms have been suggested, for example WARP, SARP, and GARP [Jehle and Reny, 2011, pp.92-96].

[11] It may be objected that *consumption* can be giving away, or that *goods* can include charitable acts. However, these interpretations do not make sense within consumer theory which only considers one isolated consumer.

Theory of the firm

Where consumer theory considers the demand side of a market, the theory of the firm considers production and the supply side of a market. Since my focus here is on the different variants of the rationality assumption, I will not go through the entire theory of the firm, but merely focus on the aspects relevant for us.

The unit considered on the supply side in an economy is called a *firm*. Firms are modelled as "black boxes" that transform input to output [Mas-Colell et al., 1995, p.127]. It is assumed that each firm is a perfect competitor on its input and output market, such that firms will be price takers (the price of input and output cannot be affected by the firm's actions):

> While the assumption of price-taking behaviour and the conditions leading to it are extreme, they provide a tractable model of the firm that is capable of yielding important insights. The competitive firm therefore merits our careful study [Jehle and Reny, 2011, p.145].

The theory of the firm depends on *the firm's problem* or *the profit maximisation problem*:

$$\max py - \mathbf{W} \cdot \mathbf{x} \quad \text{s.t.} \quad f(\mathbf{x}) \leq y \tag{5.2}$$

Here, py is the revenue from the firm's production given the produced amount y and the price of one unit of y, p. $\mathbf{W}$ is the price vector corresponding to the input vector $\mathbf{x}$, and $f(\mathbf{x})$ is the production function. Jehle and Reny [2011, pp.125-126] defend the assumption of profit maximisation with the following arguments:

> Why would someone go to the considerable bother of creating a firm in the first place, and what guides such a person in the myriad decisions that must be made in the course of the firm's activities? Profit maximisation is the most common answer an economist will give, and it is an eminently reasonable one. [...] These people are also consumers, and consumers get their satisfaction from the goods and services their income will buy. Clearly, the more profit the firm can make, the greater will be its owners' command over goods and services. In this view, firm owners will insist that all

> decisions on acquiring and combining inputs, and on marketing output, must serve the goal of maximising profit.
>
> Of course, profit maximisation may not be the only motive behind firm behaviour, and economists have considered many others. Sales, market share, or even prestige maximisation are also possibilities. Each of these alternatives to profit maximisation - and others too - have at least some superficial appeal. Yet the majority of economists continue to embrace the hypothesis of profit maximisation most of the time in most of their work.
>
> There are good reasons for this tenacity. From an empirical point of view, assuming firms profit maximise leads to predictions of firm behaviour which are time and again borne out by the evidence. From a theoretical point of view, there is first virtue of simplicity and consistency with the hypothesis of self-interested utility maximisation on the part of consumers. Also, many alternative hypotheses, [...], may be better viewed as short-run tactics in a long-run, profit-maximising strategy, rather than as ultimate objectives themselves. Finally, there are identifiable market forces that *coerce* the firm towards profit maximisation even if its owners or managers are not themselves innately inclined in that direction. Suppose that some firm did not maximise profit. Then the fault lies with the managers, and if at least a working majority of the firm's owners are non-satiated consumers, those owners have a clear common interest in ridding themselves of that management and replacing it with a profit-maximising one. If the fault lies with the owners, then there is an obvious incentive for any non-satiated entrepreneur outside the firm to acquire it and change its ways.
>
> Like the hypothesis of utility maximisation for consumers, profit maximisation is the single most robust and compelling assumption we can make as we begin to examine and ultimately predict firm behaviour. In any choice the firm must make, we therefore will always suppose its decision is guided by the objective of profit maximisation.

Thus, Jehle and Reny [2011] present three arguments for the profit maximisation assumption. The first is an empirical claim that this is what companies actually do. The second is a theoretical argument due to the connection between this assumption and the assumption presented in consumer theory. Finally, they argue that there are mechanisms that *coerce* the company to

become profit maximising. Notice that this final argument depends on either the owner *or* the manager of the company satisfying the assumptions from consumer theory, that not satisfying these assumptions is described as a *fault*, and that it is assumed that an owner - who does not satisfy the assumptions - will be willing to sell their company to an entrepreneur who does.

Varian [2014, p.347] presents a similar but shorter argument for why we should accept profit maximisation. Mas-Colell et al. [1995, pp.152-154] also argue that an owner satisfying the assumptions from consumer theory will want to maximise profit. Further, they argue that the same will be the case if the company has several owners:

> Fortunately, it is possible to resolve these issues and give a sound theoretical grounding to the objective of profit maximization. We shall now show that under reasonable assumptions this is the goal that all owners would agree upon [Mas-Colell et al., 1995, p.152].

The assumptions are that i) all owners satisfy the assumptions stated in consumer theory, ii) prices are fixed and do not depend upon actions of the firm, iii) profits are not uncertain, and iv) managers can be controlled by owners.

Thus, the variant of the rationality assumption present in the theory of the firm states that firms seek to maximise their own profit given their production function. Firms profit maximise since their owners are assumed to want to increase their incomes so that they can pay for more goods and thus increase their utilities. There are no other concerns relevant for the firm.

5.3.2 Choice under uncertainty, game theory, and social choice

Besides consumer theory - in which consumers choose between goods - economics textbooks typically present three other frameworks for making decisions. These are choice under uncertainty, game theory, and social choice theory. I will consider each in turn.

Choice under uncertainty

Choice under uncertainty is based on the theory developed by Von Neumann and Morgenstern [1944a] (VNM). Instead of considering consumption bundles, the alternatives are now *lotteries* or *gambles* defined by an objective probability distribution over a set of outcomes which are mutually exclusive. Further, we can distinguish between *simple* gambles as the one just described and *compound* gambles, where the possible outcomes of a gamble are new gambles [Jehle and Reny, 2011, p.99]. The preference relation used in choice under uncertainty is also assumed to be rational by satisfying *completeness* and *transitivity*:

> It should be emphasized that, if anything, the rationality assumption is stronger here than in the theory of choice under certainty discussed in Chapter 1. The more complex the alternatives, the heavier the burden carried by the rationality postulates. In fact, their realism in an uncertainty context has been much debated. However, because we want to concentrate on the properties that are specific to uncertainty, we do not question the rationality assumption further here [Mas-Colell et al., 1995, p.171].

Further, the preference relation is assumed to be *continuous* on the space of simple lotteries [Mas-Colell et al., 1995, p.171] and *monotone*:

> Although most people will usually prefer gambles that give better outcomes higher probability, as monotonicity requires, it need not always be so. For example, to a safari hunter, death may be the worst outcome of an outing, yet the *possibility* of death adds to the excitement of the venture. An outing with a small probability of death would then be preferred to one with zero probability [Jehle and Reny, 2011, p.100].

Finally, the preference relation is assumed to satisfy the *independence axiom*. This axiom states that if two lotteries, where the first is preferred to the second, each are combined with a third lottery, then the first will still be preferred to the second [Mas-Colell et al., 1995, p.171].

Varian [2014, pp.224-226] provides the following defence of the independence axiom:

> Bygones are bygones - so what *doesn't* happen shouldn't affect the value of consumption in the outcome that *does* happen.
>
> Note that this is an *assumption* about an individual's preferences. It may be violated. When people are considering a choice between two things, the amount of a third thing they have typically matters. The choice between coffee and tea may well depend on how much cream you have. But this is because you consume coffee together with cream. If you considered a choice where you rolled a die and got either coffee, *or* tea, *or* cream, then the amount of cream that you might get shouldn't affect your preferences between coffee and tea. Why? Because you are either getting one thing or the other: if you end up with cream, the fact that you might have gotten either coffee or tea is irrelevant.

Though the title of this subsection is *why expected utility is reasonable*, Varian gives no (further) reasons why expected utility is a good way to present choice problems under uncertainty.

In Jehle and Reny [2011, pp.100-102] the independence axiom is presented as two separate axioms. The first, *substitution axiom*, states that a person, who is indifferent between the outcomes of two gambles, and where the outcomes occur with the same probability in the two gambles, will be indifferent between the two gambles. This axiom is not discussed. The second axiom, *reduction to simple gambles*, assumes that people are indifferent between a combined gamble and its corresponding simple gamble:

> As plausible as [this axiom] may seem, it does restrict the domain of our analysis. In particular, this would not be an appropriate assumption to maintain if one wished to model the behaviour of vacationers in Las Vegas. They would probably not be indifferent between playing the slot machines many times during their stay and taking the single once and for all gamble defined by the effective probabilities over winnings and losses. On the other hand, many decisions under uncertainty are undertaken outside of Las Vegas, and for many of these, [the axiom] is reasonable [Jehle and Reny, 2011, pp.101-102].

If we assume that a preference relation on the space of lotteries satisfies all the above axioms, the *expected utility theorem* states that the preference relation can be represented by a utility function that has the *expected utility*

property such that the utility of a lottery will be the sum of the utilities for each outcome in the corresponding simple lottery times the probabilities of each of them occurring [Jehle and Reny, 2011, 102]. A utility function that satisfies the expected utility property is called a VNM utility function and is unique up to a positive affine transformation [Jehle and Reny, 2011, p.108]. Finally, it is assumed that agents are expected utility maximisers [Jehle and Reny, 2011, pp.103,106].[12]

Mas-Colell et al. [1995, pp.178-183] is the only textbook of the three that discusses expected utility theory further:

> A first advantage of the expected utility theorem is technical: It is extremely convenient analytically. This, more than anything else, probably accounts for its pervasive use in economics. It is very easy to work with expected utility and very difficult to do without it. [...]
>
> A second advantage of the theorem is normative: Expected utility may provide a valuable guide to action. People often find it hard to think systematically about risky alternatives. But if an individual believes that his choices should satisfy the axioms on which the theorem is based (notably, the independence axiom), then the theorem can be used as a guide in his decision process. [...]
>
> As a descriptive theory, however, the expected utility theorem (and, by implication, its central assumption, the independence axiom), is not without difficulties. Examples 6.B.3 [*The Allais Paradox*] and 6.B.3 [*Machina's paradox*] are designed to test its plausibility. [...]
>
> Because of the phenomena illustrated in the previous two examples, the search for a useful theory of choice under uncertainty that does not rely on the independence axiom has been an active area of research [...]. Nevertheless, the use of the expected utility theorem is pervasive in economics.
>
> Finally, one must use some caution in applying the expected utility theorem because in many practical situations the final outcomes of uncertainty are influenced by actions taken by individuals. Often these actions should be explicitly modeled but are

[12] Economics textbooks will normally also consider risk aversion. Due to space constraints I have chosen not to consider this theory here.

> not. Example 6.B.4 [*Induced preferences*] illustrates the difficulty involved. [...]
>
> Although it is not a contradiction to the postulates of expected utility theory, and therefore it is not a serious conceptual difficulty, the induced preferences example nonetheless raises a practical difficulty in the use of the theory. The example illustrates the fact that, in applications, many economic situations do not fit the pure framework of expected utility. Preferences are almost always, to some extent, induced.

In summary, the microeconomics textbook account of choice under uncertainty assumes that people aim to maximise their own expected utility. While the theory has several constraining and unrealistic assumptions, the theory itself does not require people's preferences to be self-interested. Indeed, it is possible for a person's utility to be maximised by doing well for others. However, looking at the textbooks' presentations of the theory, we notice two things. First, the decision maker's preferences towards other people are not a part of the theory. Thus, nothing in the theory suggests that other-regarding preferences should be a concern. Second - and most importantly - the examples used in the textbooks *only* consider people with self-interested preferences (see for example the quotes above). Thus, even if expected utility theory does not require people to be self-interested, the textbook presentations of it strongly suggest that they are.

Game theory

Game theory has already been considered in chapters 1 and 3. Here, I will therefore only discuss game theory's assumptions about human behaviour and how they are presented in the textbooks.[13]

In game theory, the rational player still aims at maximising their own utility or payoff given other players' choices [Jehle and Reny, 2011, pp.306,308]. It is further assumed that

a. every player is rational and all players know this [Mas-Colell et al., 1995, p.239],

[13] I will not consider behavioural economics (where real people's behaviour and choices are being studied) since it is not a part of the standard microeconomics curriculum.

b. the payoff functions of all players take an expected utility form [Mas-Colell et al., 1995, p.220],

c. all players know the structure of the game (all possible strategies for all players) and this is common knowledge [Mas-Colell et al., 1995, p.226],

d. if the game has multiple steps, each player knows everything that happened in previous steps [Mas-Colell et al., 1995, p.222],

e. every player knows the payoffs for all the other players [Jehle and Reny, 2011, p.319],

f. players have *perfect recall* so that they will never forget what they once knew, including their own actions [Mas-Colell et al., 1995, p.224].

A strategy that satisfies assumptions a. and b. is called a *rationalizable strategy*, since it is a strategy that a rational player can *justify* or *rationalize* when assumption c. is satisfied [Mas-Colell et al., 1995, p.243]. If a game does not satisfy assumptions d. and e., it is a game with *incomplete information* [Jehle and Reny, 2011, pp.319-320].

There is no discussion of the plausibility or reasonableness of the above assumptions in the three textbooks (except for the reasons why we also want to consider games of incomplete information). However, Mas-Colell et al. [1995, pp.248-249] provide five arguments why a *Nash equilibrium* is a good solution concept.[14] When considering backwards induction, Jehle and Reny [2011, p.335] provide the following discussion for the solution $(0, 0)$ to the game depicted in figure 5.1:

> It may seem a little odd that the solution to this game yields each player a payoff of zero when it is possible for each to derive a payoff of 3 by playing 'right' whenever possible. However, it would surely be a mistake for player 2 to play r' if node z is reached because player 1 will rationally choose L'' at y, not R'', because the former gives player 1 a higher payoff. Thus, player 2, correctly anticipating this, does best to choose l', for this yields player 2 a payoff zero, which surpasses the alternative of -1.

[14] They are critical towards at least two of the arguments.

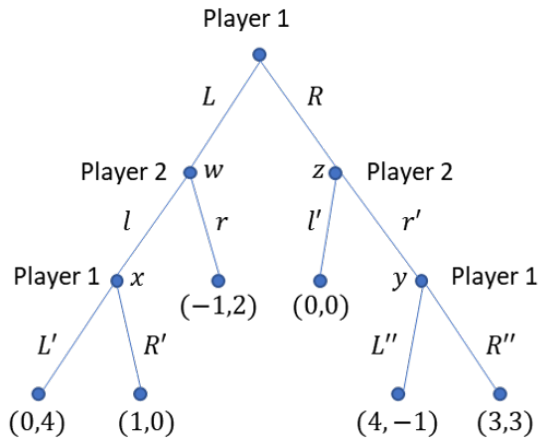


Figure 5.1: **Game example used by Jehle and Reny [2011].** An example of a game with players 1 and 2 that can be solved by backwards induction. See Jehle and Reny [2011, p.335].

In summary, the mathematical and behavioural assumptions developed in previous areas of microeconomics provide strong mathematical results in game theory. Though the theory is consistent with people's preferences being other-regarding, it is noteworthy that all examples in the three textbooks use people with self-interested preferences. Further, several examples in the books use monetary payoffs, thus assuming - as in consumer theory - that an increase in money is equivalent to an increase in utility.

Social choice theory

Social choice theory is concerned with *collective decisions*. It is a part of *welfare economics* which addresses the distribution of welfare [Varian, 2014, p.631]. Jehle and Reny [2011, p.267] introduce the topic with the following remark:[15]

> With only few exceptions, we have so far tended to concentrate on questions of 'positive economics'. We have primarily been content to make assumptions about agents' motivations and circumstances, and deduce from these the consequences of their individual and collective actions. In essence, we have characterised

[15] As discussed in chapter 4, the distinction between positive and normative economics can be traced back to Friedman [1953].

> and predicted behaviour, rather than judged it or prescribed it in any way. In most of this chapter, we change our perspective from positive to normative, and take a look at some important issues in welfare economics. At the end of the chapter we return to positive economics and consider how individuals motivated by self-interest make the problem of social choice doubly difficult.
>
> [...] Welfare economics helps to inform the debate on social issues by forcing us to confront the ethical premises underlying our arguments as well as helping us to see their logical implications.

The starting point for social choice is a group of individuals who each have a preference ordering over the same set of alternatives. It is assumed that each individual's preference relation is *complete*, *transitive*, and *reflexive*. A *social welfare functional* is a function that takes the tuple of individual preference orderings as input and creates as its output one preference ordering which is also *complete*, *transitive*, and *reflexive* [Mas-Colell et al., 1995, p.793]. A *social choice function* takes the same input but has as its output the most preferred alternative only [Jehle and Reny, 2011, p.270]. Given the further requirements of *unrestricted domain*, *weak Pareto principle*, *independence of irrelevant alternatives*, and *non-dictatorship*. Arrow's impossibility result shows that no social welfare functional will satisfy the above requirements [Jehle and Reny, 2011, pp.271-272]. This impossibility result can be avoided via the "escape routes" of *restricted domain* and *single-peakedness* [Mas-Colell et al., 1995, pp.799-806]. Finally, the *Gibbard-Satterthwaite theorem* states that is it is impossible to have a social choice function that is *strategy-proof* given some reasonable conditions.

Another way to escape the impossibility results is to move away from ordinal preference orderings such that the *intensity of preferences* can be compared across individuals:

> Before merely pushing forward, a warning is in order. The idea that 'intensity of preferences' can be compared in a coherent way across individuals is controversial at best. Nonetheless, the alternative approach to social choice that we are about to explore takes as a starting point - as an assumption - that such comparisons can be made in a meaningful way. We shall not attempt to justify

> this assumption. Let us just see what it can do for us [Jehle and Reny, 2011, pp.279-280].

Varian [2014, p.634] introduces interpersonal utility comparisons without commenting on it. Mas-Colell et al. [1995, p.818] raise some questions regarding interpersonal utility comparisons but postpone answering them with the comment: "For current purposes there is no need to answer them". Unfortunately, the section they refer to for a discussion, does not provide one [Mas-Colell et al., 1995, pp.831-838].

By allowing for interpersonal utility comparisons, it is possible to construct a *social welfare function* that has cardinal utility functions as its input and output. Examples of such social welfare functions are the *utilitarian* where the outcome is the sum of the individuals' utilities, and the *"Rawlsian"* or *maxi-min* where the outcome is the utility of the least well off. Given these functions and a *utility possibility set*, the *problem of welfare maximisation* is to determine how to maximise the utility of a group of people given the social welfare function and constrained by the goods available [Varian, 2014, p.636]. As already discussed in section 3.3, Varian [2014] and Mas-Colell et al. [1995] abstain from judging between the different welfare functions while Jehle and Reny [2011, pp.284,288-290] suggest that it is reasonable to use a utilitarian one.

Summing up, social choice theory assumes that individuals want to maximise their own preference ordering which satisfy the now standard requirements. The theory is compatible with people having other-regarding preferences or preferences interpreted as judgements about the common good. Unfortunately, this point is not mentioned in any of the textbooks. Using cardinal utility functions which enable interpersonal utility comparisons, it is assumed that *society* wants to maximise utility given a social welfare function and its utility possibility set. This can be seen as a "societal" variant of the rationality assumption.

5.3.3 Market behaviour and general equilibrium

Finally, I present the two standard theories used to analyse market behaviour in microeconomics. The first is *competitive markets* or *partial equilibrium*, where a market for one good is considered. The second is *general equilibrium*, where several markets (each for one good) are analysed together. Both theories use the variants of the rationality assumption developed in consumer theory and theory of the firm.

Competitive markets

Mas-Colell et al. [1995] introduce the topic *market equilibrium and market failure*, as follows:

> our focus shifts to the fundamental issue of economics: *the organization of production and the allocation of the resulting commodities among consumers*. This fundamental issue can be addressed from two perspectives, one *positive* and the other *normative*.
>
> From a positive (or *descriptive*) perspective, we can investigate the determination of production and consumption under various institutional mechanisms. The institutional arrangement that is our central focus is that of a *market* (or *private ownership*) *economy*. In a market economy, individual consumers have ownership rights to various assets or goods. Likewise, firms, which are themselves owned by consumers, decide on their production plan and trade in the market to secure necessary inputs and sell the resulting outputs. Roughly speaking, we can identify a *market equilibrium* as an outcome of a market economy in which each agent in the economy (i.e., each consumer and firm) is doing as well as he can given the actions of all other agents.
>
> In contrast, from a normative (or *prescriptive*) perspective, we can ask what constitutes a *socially optimal* plan of production and consumption (of course, we will need to be more specific about what "socially optimal" means), and we can then examine the extent to which specific institutions, such as a market economy, perform well in this regard [Mas-Colell et al., 1995, p.307].

The market studied is a *competitive market* of one good. In this market, the price of the good is publicly known and all agents are price takers [Mas-Colell et al., 1995, p.307]. It is assumed that agents have perfect knowledge, that

each firm will profit maximise and that each consumer will utility maximise as in consumer theory - but where their income now depends on the price of the good. Thus, the assumptions regarding consumer and firm behaviour

> reflect the underlying assumption, common to nearly all economic models, that agents in the economy seek to do as well as they can for themselves [Mas-Colell et al., 1995, p.314].

Further, the market will *clear* such that the aggregated supply of the good will equal the aggregated demand for the good at the equilibrium price [Mas-Colell et al., 1995, pp.314-315].

In the short run, the given number of firms in the market is fixed [Jehle and Reny, 2011, p.166]. However, in the long run, there is *free entry and exit* so that firms will enter if they will be able to gain a positive profit and exit if they get a negative profit [Mas-Colell et al., 1995, p.335]. According to Mas-Colell et al. [1995, p.334], this "is often a reasonable approximation when we think of long-run outcomes in a market". Only Varian [2014, p.416] mentions that some industries may have *barriers to entry*. However, he does not discuss them any further. In a long run equilibrium, the market is assumed to clear, the profit of each firm will be zero, and no firm will have incentive to exit or enter [Jehle and Reny, 2011, p.168].

The outcome of a market equilibrium is evaluated by introducing the concept of *Pareto efficiency* which has already been discussed. Mas-Colell et al. [1995, p.313] provides the following discussion of the principle:

> It is important to note that the criterion of Pareto optimality does not insure that an allocation is in any sense equitable. For example, using all of society's resources and technological capabilities to make a single consumer as well off as possible, subject to all other consumers receiving a subsistence level of utility, results in an allocation that is Pareto optimal but not in one that is very desirable on distributional grounds. Nevertheless, Pareto optimality serves as an important minimal test for the desirability of an allocation; it does, at the very least, say that there is no waste in the allocation of resources in society.

According to the first fundamental theorem of welfare economics, a market equilibrium in the markets just discussed will be Pareto efficient [Mas-Colell et al., 1995, p.326].

In Varian [2014, ch.34,36] and Mas-Colell et al. [1995, ch.11] some possible imperfections of the above market are discussed; namely the cases of *externalities* and *public goods*. Jehle and Reny [2011] do not mention either of these considerations. Here, I will shortly focus on public goods since they give rise to the *free rider problem*. A public good is a good where one agent using it does not preclude other agents from also using it [Mas-Colell et al., 1995, p.359]. Examples are TV channels, firework shows, or public parks. Public goods are problematic since it can be difficult to *exclude* people from using them. Thus, the free-rider problem occurs: it is in everybody's self-interest that they are there, but it is also in everybody's interest to let other agents pay for them [Mas-Colell et al., 1995, p.362]. In order to solve this problem, Mas-Colell et al. [1995, p.362] suggest:

> The inefficiency of private provision is often remedied by governmental intervention in the provision of public goods. Just as externalities, this can happen not only through quantity-based intervention (such as direct governmental provision) but also through "price-based" intervention in the form of taxes or subsidies.

This appeal to governmental intervention is not found in Varian [2014, pp.711-715], who proposes a certain form of collective auction initiative instead. It is however interesting, since both books earlier have stressed the inefficiency or *deadweight loss* of any form for taxation in competitive markets.

With regard to taxation, Varian [2014, p.304-306] writes:

> The lost output is the social cost of the tax. [...] The government *gains* revenue from the tax. And, of course, the consumers who benefit from the government services provided with these tax revenues also gain from the tax. We can't really say how much they gain until we know what the tax revenues will be spent on. [...]
>
> This area [the lost output] is known as the **deadweight loss** of the tax or the **excess burden** of the tax. This latter phrase

> is especially descriptive. [...] What is the source of this excess burden? Basically it is the lost value to the consumers and producers due to the reduction in the sales of the good. You can't tax what isn't there. [footnote: At least the government hasn't figured out how to do this yet. But they're working on it.] So the government doesn't get any revenue on the reduction in sales of the good. From the viewpoint of society, it is a pure loss - a deadweight loss.

Varian's argument is interesting since he only looks at the monetary cost of a tax. Thus, he does not consider how much utility is lost by this monetary cost compared to how much utility is gained - for example by redistributing the money to people with low income or to public goods. Since money also have marginal decreasing utility, there might very well be a gain in utility even though there is a decrease in production. Mas-Colell et al. [1995, pp.331-334] consider welfare directly when discussing taxation. Their language is similar to that of Varian, though more mathematics is introduced. However, Mas-Colell et al. [1995, p.334] end their discussion by noting that the analysis only applies as long as its assumptions are satisfied. If this is not the case, redistribution can lead to a potential Pareto improvement.

General equilibrium theory

General equilibrium theory is used to determine equilibrium prices and quantities for multiple goods in a perfectly competitive market system [Mas-Colell et al., 1995, p.511]. Jehle and Reny [2011, p.195] link the theory to "[Adam] Smith's vision of a smoothly functioning system composed of many self-interested individuals buying and selling on impersonal markets - with no regard for anything but their personal gain" and introduce it with:

> we do not merely ask under what conditions a set of market-clearing prices exists. We also ask how well a market system solves the basic economic problem of distribution. We will begin by exploring the distribution problem in very general terms, then proceed to consider the existence of general competitive equilibrium itself. Along the way, we will focus particular scrutiny on Smith's claim that a competitive market system promotes

> society's welfare through no conscious collective intention of its members.

In order to tackle the problem of general equilibrium theory, microeconomics textbooks start with a simplified *barter economy*.[16] Here, consumers are assumed to have a fixed endowment of each good so that it can be studied how they will trade among themselves in order to maximise their utilities [Varian, 2014, pp.582-583]. It is further required that each consumer is a price taker and knows everything about their own and other consumers' preferences and bundles of goods. Given this, it can be shown that all resulting allocations of goods in the barter economy are Pareto efficient [Jehle and Reny, 2011, pp.200,202].[17]

Moving to a market economy, this is assumed to be a closed and interrelated system where the values of the equilibrium quantities and prices are determined simultaneously for all goods in the markets [Mas-Colell et al., 1995, p.511]. It is further assumed that consumers and firms are price takers, that firms aim to maximise their profits, and consumers aim to maximise their utilities given their endowments. Each consumer knows all the goods sold in all markets along with their prices, and they can be fully confident that there is sufficient supply to satisfy the bundle they want to purchase independently of what other consumers demand. Similarly, each firm knows the inputs, outputs, and prices of all goods in all markets, and it can be fully confident that it can produce the amount of goods that maximises its profit without worrying about how much other firms produce [Jehle and Reny, 2011, pp.201-202]. Here, Jehle and Reny [2011, p.202] write:

> The naivete expressed in the decentralised aspect of the competitive model (i.e., that every agent acts in his own self-interest while ignoring the actions of others) should be viewed as a strength. Because in equilibrium consumers' demands *will* be satisfied, and because producers' outputs *will* be purchased, the actions of the

[16] This is introduced by an *Edgeworth box* illustrating two people exchanging two goods.

[17] Varian [2014, p.588] states that the analysis might be implausible if there are only two people in the barter economy. Instead, we can consider two *types* of consumers, with many consumers of each type. The analysis "makes perfect sense in the many-person case, which is what we are really concerned with".

> other agents *can* be ignored and the *only* information required by consumers and producers is the *prevailing prices*. Consequently, the informational requirements of this model are minimal.

Finally, no price must be zero (since this will create infinite demand), the *aggregate excess demand* should be possible to represent by a function that satisfies the requirements of *continuity* and *homogeneity*, and it should always be zero at any set of positive prices [Jehle and Reny, 2011, p.204].

Given these assumptions, it can be shown that there exists at least one equilibrium [Mas-Colell et al., 1995, pp.585]. Further, the first fundamental theorem of welfare economics states that the equilibrium will be Pareto efficient, and the second fundamental theorem of welfare economics states that we can reach any Pareto efficient allocation (which will correspond to an equilibrium in the market system) by changing the initial endowments of the consumers and let the market do the rest [Jehle and Reny, 2011, pp.217-220].[18] While the second theorem is not discussed in Jehle and Reny [2011], Varian [2014, pp.604-606] argues that the theorem can be used to support a tax on initial *endowment*. However, this endowment is people's endowment of labour - how much they *could* work - rather than how much they *choose* to work. Here, Mas-Colell et al. [1995, pp.556-557] and Varian agree that it will be very difficult to tax in reality, and Varian [2014, p.605] concludes that "no one is advocating such a radical restructuring of the tax system".

Finally, the importance of perfect information is show by discussing how a market can fail under asymmetric information. Given specific utility, demand, and supply functions, we can end up in a situation where no transaction will occur - despite there being sellers who want to sell at a price where there are buyers who want to buy [Akerlof, 1970].[19]

[18] The theorem depends on the assumption that consumer preferences and firms' production sets are convex [Mas-Colell et al., 1995, p.308].

[19] See Mas-Colell et al. [1995, p.437], Varian [2014, p.719], and Jehle and Reny [2011, pp.382,416].

5.4 Discussion: the strength and persistence of the rationality assumption

The review of microeconomics textbooks in section 5.3 shows that the standard topics considered in microeconomics all rely on some variant of the rationality assumption. Looking at the text excerpts, we further see that the most used justification of the assumption is a version of:

- It is reasonable, hardly objectionable, natural, or almost tautological.

The second most used justifications are:

- It is necessary for economic theory or consistent with other parts of economics.
- It gives us a simple model that is easy to work with and can provide insights.

The third most used justifications (or lack thereof) are:

- We ought to act this way, or the theory can be used as a guide for action.
- Most economists assume it.
- We will not question the assumption.
- It is typically the case, though there are a few counter examples.

Finally, we have one time encountered the justification that:

- There is empirical evidence that it is the case.

As also indicated in section 5.3, however, several of the requirements in the rationality assumption are difficult to defend as "natural" or in any way plausible for normal human beings, and many of them can be questioned or disputed. It is therefore relevant to consider *why* the rationality assumption is so prominent in neoclassical microeconomics.

First, as we saw in section 5.2, the development of the rationality assumption is closely related to the development of economic theory as it is presented in textbooks today [Giocoli, 2003, p.3]. Without the changed description of

human behaviour, Mill would not have been able to distinguish economics from other moral sciences, and Jevons would not have been able to translate economic problems to the language of mathematics. Finally, Menger's focus on choice behaviour - present in the requirement of rational preferences and revealed preference theory - helped give economics the credibility of a science studying observable phenomena. Thus, one reason the rationality assumption is so persistent is that its formulation played an important role in the creation of neoclassical microeconomics as we know it today.

A second reason is that economic models (like the ones described in 5.3) are subject to rules that determine what can be done with them. The rationality assumption and the assumption of an equilibrium tendency are prominent in economics because they are the only rules that are both *formal* and *economic* [Morgan, 2012, pp.394-395]. *Formal rules* are determined by the language, the model is "written" in: a mathematical model has different constraints than a physical model made of water pipes and it is possible to do different things with them. The rationality assumption is a formal rule because its idealisations and exaggerations of human behaviour makes it possible to consider how agents in a mathematical model will behave. Looking at the textbook versions of the assumption, we see that several of its requirements - like continuity, convexity, and monotonicity - are assumed only because they increase the mathematical ease of using the assumption. *Economic rules* constrain and determine how a model can be manipulated on the basis of "the economic subject matter represented in the model" [Morgan, 2012, p.26]. An example of this kind of rule is the interpretation of the numbers depicted in a game matrix as utilities. The rationality assumption is an economic rule because it defines how agents will act in the model with reference to how people (ideally) are motivated to act in economic situations. This discussion of how people act in economic situations is present both in the historical and current accounts of human behaviour in economics and is used - for example - to explain why trade takes place. Because the rationality assumption is both a formal and economic rule, it has - together with the equilibrium assumption - become the most important assumption in neoclassical economics. Indeed, it is necessary (though never sufficient) for a

model in economics to use at least one of these assumptions [Morgan, 2012, p.395]. In microeconomics - considering the actions of individual agents - the prominent assumption is the rationality assumption. In this way, the assumption binds microeconomic theory together, even though its models consider different subject matters.

The above reasons explain why the rationality assumption is such a strong and persistent assumption in neoclassical microeconomics. However, for the purpose of this thesis, it is also relevant to consider *how* the assumption is presented in textbooks today.

Recall the discussion of Friedman's methodology from chapter 4. Besides distinguishing between positive and normative economics, Friedman [1953, p.153] argues that a theory should not be judged by the truth of its assumptions, but only by the correctness of its predictions. This is because model assumptions are always descriptively false. Indeed, economic models are not supposed to produce an exact pictures of the world, but rather to make a small engine that can analyse what will happen in the world [Friedman, 1953, p.167]. If the engine works, then the truth or falsity of its assumptions is of no consequence. As an example, Friedman [1953, p.158] argues that the rationality assumption (that people aim at maximising their monetary gain) is an acceptable assumption to have because the models it is used in make accurate predictions. Thus, Friedman's methodological instrumentalism defend the rationality assumption despite it being descriptively false. Looking at the summarised justifications of the assumption in the textbooks, however, we see that it is described as natural, and as the way we ought to act. Thus, the textbooks present the assumption both as *descriptively plausible* and *normatively desirable*. This is contrary to methodological instrumentalism and to Friedman's distinction between positive and normative economics. As I will argue in the next chapter, it is also a part of the mechanism that makes microeconomics self-fulfilling.

Given the findings from this and the previous chapter that Friedman's distinction between positive and normative economics is not satisfied historically nor in contemporary microeconomics textbooks, it is relevant to question whether the distinction - though theoretically possible - is desirable,

or even attainable, to maintain in practice. Looking at the similar debate within theory choice on non-epistemic values in science, there is compelling evidence that science is not value free and that trying to maintain an ideal of value neutral science can have harmful consequences [Kuhn, 1977, Longino, 1990, 1994, Lacey, 2004, Anderson, 2004, Wylie and Nelson, 2007]. Thus, I suggest that the distinction is problematic to use with regard to microeconomic theorising.

Finally, though some variants of the rationality assumption discussed in section 5.3 can include other-regarding preferences, it is important to note that this is not mentioned in any of the textbooks when introducing the models.[20] Instead, the textbooks repeatedly stress that people act in accordance with their self-interest. Further, when *gains* are defined as *preferences* or *utilities*, the textbooks still provide examples where people maximise their *incomes*. Thus, the textbooks assume (without discussion) that in economic situations involving money, people will always satisfy their self-interest by satisfying their narrow self-interest. This relation between self-interest and narrow self-interest can be seen as originating in consumer theory where it is assumed. Since consumer theory - despite its very narrow scope - is a building block for the theory of the firm, competitive markets, and general equilibrium theory, the relation is also satisfied here. This - combined with the long standing historical tradition of promoting the benefits of (narrow) self-interest in economics (discussed in chapter 4) - may be the reason why the textbooks also transfer the assumption of (narrow) self-interest to choice under uncertainty and game theory, despite the fact that it is not required for these theories.

[20] This analysis holds for the standard microeconomics models discussed in section 5.3. Notice, that some of the textbooks do include theories of collaboration, e.g. Mas-Colell et al. [1995, pp.673-684,825-849]. Such theories are, however, kept in the final chapters or as an appendix in the textbooks. It is therefore doubtful that they will be included in any traditional microeconomics course.

5.5 Conclusion

In this chapter, I have provided the final background analyses needed for the main argument in the thesis. I did this by addressing the questions *why* the rationality assumption is so prominent in microeconomics textbooks today as well as *how* it is presented in these textbooks.

By analysing the historical development of the rationality assumption - exemplified by the works of Smith, Malthus, Mill, Menger, Jevons, and Knight - we see how the development of the assumption is closely related to the development of economic theory. Thus, one reason the rationality assumption is so prominent today is that it played a crucial role in differentiating economics from other science and in transforming economics from a verbal to a model-based, formal science. The strength and persistence of the assumption is further supported by it being both a formal and an economic rule.

Looking at *how* the rationality assumption is presented in contemporary microeconomics textbooks, we see that a variant of the assumption is present in all standard textbook models. Further, the text excerpts show that - contrary to Friedman's methodologies - the textbooks justify the rationality assumption as being both *descriptively plausible* and *normatively desirable*. Finally, the excerpts show that despite the fact that some variants of the rationality assumption do not require that people act in accordance with their (narrow) self-interest, the textbooks do not account for this fact. Instead, their examples are explicitly concerned with self-interested people, who are typically also narrowly self-interested.

Given these findings, I now turn to the main argument of the thesis. In chapter 6, I present a theoretical argument for the claim that microeconomics can be self-fulfilling by promoting a social norm of (narrow) self-interest in economic situations through the dissemination of microeconomics textbooks and teaching practices. In chapter 7, I support the argument with empirical results from a series of experiments designed to test it.

Part II

HOW MICROECONOMICS CAN BECOME SELF-FULFILLING

Chapter 6

Promoting a social norm of (narrow) self-interest

6.1 Introduction

Summarising part one of the thesis, I first proposed to use *self-fulfilling science* to describe cases where the dissemination of scientific descriptions concerning all agents can have social implications by influencing behaviour. Next, I argued that the current empirical literature shows that microeconomics is self-fulfilling in economic situations due to the dissemination of the rationality assumption as it is presented in the standard textbooks models. Further, the empirical literature shows that in non-economic situations, or in situations where there are salient non-economic priorities (such as promise-keeping), the self-fulfilling effects of microeconomics are reduced. In order to prepare the argument for *one way* in which microeconomics can be self-fulfilling, I then turned to a historical analysis of economic theory, showing how the normative claim that it is beneficial to act self-interestedly and greedy has been a part of modern economic theory from the beginning and still is today. The analysis also shows that Friedman's argument that positive economics is independent of normative judgements does not hold for neoclassical economics, and that

the arguments defending greed and self-interest fail to do so convincingly. This prompts the question why the assumption that humans act (narrowly) self-interested is still present in microeconomics textbooks today. Looking at the development of the rationality assumption, we see that it is closely related to the development of economic theory in general. Further, its strength and persistence in microeconomics is supported by the fact that it is both an economic and a formal rule. Finally, I presented the current variants of the rationality assumption in the standard microeconomics models along with the informal textbook discussions of the variants. The textbook discussions show that the assumption is defended as both descriptively plausible and normatively desirable. Further, the textbooks do not account for the fact that some variants of the rationality assumption can be satisfied by agents with other-regarding preferences. Instead, they only use examples with self-interested people that are typically also narrowly self-interested.

In this part of the thesis, I use the above findings to argue that one way in which microeconomics can become self-fulfilling is by promoting a social norm of (narrow) self-interest in economic situations via the dissemination of microeconomics textbooks and teaching practices. In order to do this, I use Bicchieri's [2006 and 2016] definition of social norms and the framework surrounding it. The argument is twofold. In this chapter, I present a theoretical argument for how exposure to microeconomics textbooks and teaching practices can promote a social norm of (narrow) self-interest. I do this by combining the analyses from part I with Bicchieri's proposed mechanisms for how social norms emerge and are changed. In chapter 7, I then provide empirical evidence that people who are exposed to textbook microeconomics do change their behaviour in accordance with Bicchieri's conditions for following a social norm.

The chapter is structured as follows. In section 6.2, I introduce Bicchieri's definition of social norms and discuss two mechanisms used to explain how social norms emerge and are changed. In section 6.3, I turn to the case of microeconomics and argue how microeconomics textbooks can make their readers inclined to follow a social norm of (narrow) self-interest in economic situations. In section 6.4, I argue how microeconomics teaching practices

at universities can stabilise a social norm of (narrow) self-interest among students so that they will follow the norm and keep doing so. I end the section by arguing that the claim that microeconomics textbooks and teaching practices promote a social norm of (narrow) self-interest is supported by the empirical findings discussed in chapter 3.

6.2 Social norms and how they emerge

The aim of this section is to provide the definitions and relevant analyses needed in order to demonstrate how microeconomics textbooks and teaching practices can make people follow a social norm of (narrow) self-interest. I do this, first, by presenting Bicchieri's definition of social norms. Next, I turn to the mechanisms which Bicchieri uses to explain how social norms emerge and are changed and argue that they provide a plausible account of this. In subsection 6.2.2, I discuss the role of psychological categories and scripts, and in subsection 6.2.3 I discuss the role of reference networks. The former is relevant for the argument of how reading microeconomics textbooks can make people inclined to follow a social norm, while the latter is relevant for the argument of how microeconomics teaching practices can stabilise the behaviour recommended by the social norm.

6.2.1 *Defining social norms*

Social norms are informal rules that govern behaviour within groups and in societies [Bicchieri et al., 2018]. Such rules have been studied in anthropology, sociology, social psychology, psychology, economics, law, and philosophy and several different definitions have been suggested.[1]

[1]In social psychology, for example, Shaffer [1983, p.277] reports that three social psychology textbooks - published the same year - provided three different definitions of social norms: 1) "rules indicating what is considered to be acceptable or appropriate behaviour for the members of some group [Byrne and Baron, 1981, p.268]", 2) "a behavior form that is shared by members of a recognizable group and that may be considered to be "normal" for that group [Lindgren and Harvey, 1981, p.536]", and 3) "a widely shared expectation about appropriateness of a given behavoir in a given situation [Gergen and Gergen, 1981, p.497] ".

I have chosen to use Bicchieri's definition of social norms, first, because it is operational, so that it can be used in empirical studies. Second, it does not merely rely on observed behaviour or normative attitudes. This acknowledges that social norms can exist without the relevant behaviour ever being observed.[2] It also acknowledges that people sometimes follow social norms that they do not approve of.[3] Finally, Bicchieri's framework enables us to distinguish between different types of situations which makes it possible to use her framework for practical interventions.

Bicchieri's definition of social norms

Bicchieri [2006, p.11] defines a social norm as follows (see also definition 1):

> **Conditions for a Social Norm to Exist:**
> Let R be a *behavioral rule* for situations of type S, where S can be represented as a mixed-motive game. We say that R is a social norm in a population P if there exists a sufficiently large subset $P_{cf} \subseteq P$ such that, for each individual $i \in P_{cf}$:
>
> ***Contingency:*** i knows that a rule R exists and applies to situations of type S;
>
> ***Conditional preference:*** i prefers to conform to R in situations of type S on the condition that:
>
> **(a)** ***Empirical expectations:*** i believes that a sufficiently large subset of P conforms to R in situations of type S;
>
> and either
>
> **(b)** ***Normative expectations:*** i believes that a sufficiently large subset of P expects i to conform to R in situations of type S;
>
> or

[2] An example of this can be found in the Ik community described by Turnbull [1987]. Here, there is a strong norm of reciprocity, but because people's conditions made it difficult for them to reciprocate, they would strive never to be in the role of gift-taker. Thus, the norm of reciprocity was never observed [Bicchieri, 2006, p.9].

[3] An example of the latter case can be found in the norm of female genital mutilation (FGM) in some countries. Here, the MICS 2006 survey (table CP5) reports that the prevalence of FGM was 72.2 percent in Burkina Faso, but that only 11.1 percent of the population supported the practice [Bicchieri, 2016, p.47].

> **(b')** ***Normative expectations with sanctions:*** *i* believes that a sufficiently large subset of *P* expects *i* to conform to *R* in situations of type *S*, prefers *i* to conform, and may sanction behavior.
>
> A social norm *R* is *followed* by a population *P* if there exists a sufficiently large subset $P_f \subseteq P_{cf}$ such that, for each individual $i \in P_f$, conditions 2(a) and either 2(b) or 2(b') are met for *i* and, as a result, *i* prefers to conform to *R* in situations of type *S*.

Here, we first see that Bicchieri distinguishes between the *existence* of a social norm and a social norm being *followed*.

Let us start with the conditions for a social norm to exist. A social norm is a behavioural rule that applies to one or more *types* of situations [Bicchieri, 2006, p.12].[4] It *exists* in a *population* if each individual in a *sufficiently large* subset of the population satisfies four conditions. The requirement that the subset is sufficiently large means that everyone in a population does not need to know that a social norm exists in order for it to do so. Further, what counts as sufficiently large will vary from individual to individual [Bicchieri, 2006, p.152]. Thus, it is each individual's *belief* about the size of the subset that determines whether their threshold for sufficiently large has been reached.

In order for a behavioural rule to be a social norm, four conditions have to be satisfied. The ***contingency condition*** (condition 1) simply states that each individual in the subset knows that the behavioural rule exists and applies to this situation.

The ***conditional preference condition*** (condition 2) states that individuals in the subgroup will have a *preference* for following the rule *conditional* on the next two conditions being met [Bicchieri, 2006, p.20]. It is supported by Fehr and Schurtenberger [2018] who argue that cooperation in social dilemma experiments can be explained by a social norm of conditional cooperation.[5] The fact that the preference is conditional means that a social norm can exist in a situation without anyone following it. It also means that a person's

[4] I will disregard the requirement that it should be possible to represent the situations as mixed-motive games. For a discussion of the requirement, see subsection 6.3.1.

[5] See also Bicchieri [2006, pp.140-141].

preference for following a norm will change if the next two conditions are no longer satisfied.

The third condition is that of ***empirical expectations*** (condition 2(a)). It states that the individual should *believe* that a sufficiently large subset of the population conforms to the behavioural rule in the specific situations. It is important to stress that an individual's empirical expectation may not be the same as the actual amount of people who conform to the norm. However, we should expect that an individual will update their beliefs depending on what they observe. Thus, an individual who originally believed that everyone conformed to a behavioural rule may come to change that belief after observing sufficiently many people deviating from it in a given situation [Bicchieri, 2006, p.13].

The final condition is that of ***normative expectations (with sanctions)*** (condition 2(b)). This condition says that the individual should *believe* that sufficiently many people in the population *expect* the individual to conform with the behavioural rule in this situation. Thus, what matters for the condition to be satisfied is the individual's own beliefs about what others expect [Bicchieri, 2016, p.59]. The condition comes in two versions in order to indicate that for some individuals, the possibility of sanction - positive or negative - may be their primary reason for following a norm.[6] Note that sanctions need not actually occur. The individual just has to believe that some people *will* be willing to sanction. Here, I follow Bicchieri [2016, p.66] and merge the two versions of condition 2(b) to one:

> it is believed that sufficiently many others believe the rule should be followed, and/or may be willing to sanction deviations from it (normative expectations).

Given the above, a social norm *exists* if a sufficiently large subset of a population knows that a behavioural rule exists and applies to a specific situation, and that each individual in that subset has a conditional preference to conform to the rule in the situation, provided they believe that (a) sufficiently many others will also conform with the rule and (b) sufficiently many

[6] See for example Axelrod [1986] for an account of how the possibility of sanction can change people's behaviour.

others expect them to conform to the rule (and may be willing to sanction behaviour) [Bicchieri, 2016, p.35]. The people in this subset of the population are called *conditional followers*.

A social norm is *followed* in a population, if a sufficiently large subset of the conditional followers satisfy conditions 2(a) and 2(b), such that they prefer to conform to the rule in a given situation. Notice that the subset of followers may be smaller than the subset of conditional followers, if some conditional followers do not believe that conditions 2(a) and 2(b) are met for sufficiently many. This also means that norm followers need not expect everyone to follow a norm in order for them to do so. The important part is that sufficiently many people in the population are expected to follow it. Finally, a norm follower may stop following a social norm, if they no longer satisfies conditions 2(a) and 2(b).

I use *inclined to follow a social norm* about norm followers who know that a social norm exists and follow that norm unless they receive sufficient information to conclude that conditions 2(a) and 2(b) are not satisfied in a situation. I say that a social norm is *stabilised* if sufficiently many people continue satisfying conditions 2(a) and 2(b) in a situation so that they will follow the social norm and keep doing so.

Other types of behavioural rules

Besides social norms, Bicchieri considers four other types of behavioural rules. These are personal norms (or moral rules), collective habits, descriptive norms, and conventions. The distinctions between these rules and social norms are useful for understanding behaviour and how to mediate changes. However, it is important to realise that in real life, the boundaries between them are often blurred, and can be overlapping with regard to some actions [Bicchieri, 2006, pp38-39].

Personal norms - like refraining from killing people - are behavioural rules that we have reasons to follow independently of what other people do and that we (in principle) have a moral obligation to follow unconditionally [Bicchieri, 2006, p.20].

Collective habits - like brushing your teeth - are actions that several people do, but where the reasons for doing them are independent of this fact [Bicchieri, 2006, pp.21,31]. Thus, both personal norms and collective habits differ from social norms by not satisfying condition 2.

Descriptive norms - like driving in the left side of the road in the UK - *exist* if condition 1, condition 2, and condition 2(a) are satisfied for a sufficiently large subset of the population. Further, they are *followed* if, for a sufficiently large subset of its conditional followers, condition 2(a) is satisfied such that they will conform to the norm [Bicchieri, 2006, pp.31-32]. Thus, descriptive norms differ from social norms by not satisfying condition 2(b). Condition 2(b) is not needed for descriptive norms because it is always in one's own interest to follow them [Bicchieri, 2006, p.29]: if people in the UK drive in the left side of the road, it is in everyone's interest to keep doing this, as long as they expect everyone else to also keep doing it.[7]

Finally, *conventions* are descriptive norms where the cost of deviation from the behavioural rule is large for society [Bicchieri, 2006, pp.39-40]. Which side of the road to drive in is therefore also an example of a convention. Conventions differ from social and descriptive norms since they only exist if they are followed [Bicchieri, 2006, p.38].

6.2.2 Categorisation, schemata, and scripts

In order to understand the *cognitive mechanisms* that can underpin the emergence and change of social norms, Bicchieri [2006 and 2016] along with Bicchieri and McNally [2018] use the theories of categories, schemata, scripts, and semantic networks from psychology [Busselle, 2017, Pirnay-Dummer et al., 2012]. Here, I first present the theories as they relate to social norms. Next, I suggest that the theories provide us with a credible account of how social norms emerge and are activated. Finally, I use the theories to discuss what will happen in situations where more than one social norm is activated. Note that none of the processes described in the following have to be

[7] Bicchieri [2006, p.31] defines descriptive norms as behavioural rules that solve situations that can be represented as coordination games.

occurrent [Bicchieri, 2006, pp.84,97 and 2016, p.128].

The theory of scripts and categories applied to social norms

Recall from the definition that social norms only apply to specific types of situations. Thus, in order for a social norm to be *activated*, people have to be in a situation where they recognise that it applies [Bicchieri, 2006, p.59].

According to schema theory, we navigate the social world by *categorising* the situations we are in. A category - in this context - can be understood as a collection of situations that are similar in some way. The category can for example be represented by a prototype situation or a set of exemplars.[8] Whenever we are in a situation, we (subconsciously) try to categorise it. Since we categorise situations based on similarities, one situation may fit several different categories, depending on what is (made) salient to the person in that situation. Thus, how a situation is framed or which situational cues a person focuses on in a situation will have an effect on how that situation is categorised for them [Brewer and Treyens, 1981, Bicchieri, 2006, pp.55,58,77,86,87]. Further, the more familiar a person is with a category, the easier it will be for them to recognise or interpret a situation as belonging to that category [Nosofsky et al., 2018, Bicchieri, 2006, pp.86,88].

Once a situation has been categorised, a certain *schema* is activated. A schema is a cognitive structure that represents generic, stored knowledge about people, events, or roles [Bartlett, 1932].[9] It provides us with a theory of how a certain situation will come to pass; how certain people act, or what people with different roles are supposed to do. When a schema is activated, we tend to perceive the situation through that schema, and act or perceive people accordingly. Note that since one situation can be categorised in many different ways, it is possible that different people or even the same person can activate different schemata for one situation. This can especially occur if the situation is ambiguous [Bicchieri and McNally, 2018, pp.27,30].

[8] See for example Cantor et al. [1982], Nosofsky et al. [2018], Hampton [2016], Lech et al. [2016], Patalano et al. [2001], and Bicchieri [2006, pp.79-85].

[9] See also Nishida [2005], Mandler [2014], Bicchieri [2006, p.93], Bicchieri [2016, pp.131-132], and Bicchieri and McNally [2018, p.26].

Schemata that are used to make sense of events are called *scripts* [Abelson, 1981, Guirguis, 2003, Eaton et al., 2016]. Scripts describe appropriate sequences of stylised actions along with each actor's role in different situations [Bicchieri, 2006, pp.93,94]. As an example, consider a script for greeting a friend. Depending on your culture, you may be expected to shake hands, hug, or kiss your friend's cheeks a certain number of times.

According to Bicchieri [2006, p.94] social norms are embedded into scripts. Once we have categorised a situation as belonging to a specific class, a script will be activated which includes expectations about how to behave in the situation [Bicchieri, 2006, p.171]. Thus, for familiar types of situations, the script can provide a prior expectation of norm compliance, and people will follow this norm by default, unless something in the situation changes one's prior empirical or normative expectations. Further, Anderson [1983] found that the more people imagine themselves engaging in a certain script, the more likely they are to form the intentions to do so.

The literature suggests that schemata can be changed in at least three ways [Wicks, 1992, Bicchieri and McNally, 2018, p.35]. First, people can gradually change their schema as they observe more and more cases of divergent behaviour [Rothbart, 1981]. This is called the *bookkeeping model*. Second, a few highly significant cases of divergent behaviour can make people change their schema [Paek and Hove, 2018]. This is called the *conversion model*. Third, people might take the situations with divergent behaviour to be a subgroup of the overall schema, making a subschema for this group. This is called the *subtyping model* [Hewstone, 1994].

Finally, scripts and schemata do not exist in isolation [Bicchieri, 2006, pp.71-72]. Rather, they are embedded in what is called a *semantic network* [Collins et al., 1969]. In this network, each schema is represented as one node and the edges represent whether two schemata are related. The semantic network helps illustrate how some schemata are more related than others. Once a script or schema is activated, related scripts and schemata are also primed for activation [Colins and Loftus, 1975, Bicchieri, 2016, p.132]. These related schemata and scripts may in turn affect our understanding of the situation and come to influence our beliefs or behaviour.

Are scripts and norms related?

The above theories can be used to explain why people come to satisfy conditions 1 to 2(b) for a behavioural rule in some situations. This makes it a useful starting point for arguing how microeconomics can plausibly promote a social norm. However, in order to make this argument, it is relevant to discuss whether there is evidence that the theories can be used to explain how social norms are activated and changed.

First, it is important to clarify that the theories of categorisation, schema, scripts, schema change, and semantic network all have been supported empirically. Thus, they provide a credible account of how we understand and behave in different types of situations.[10]

Looking at the link between social norms and the above theories, no one has conducted a controlled experiment to determine how script changes relate to interventions aimed at changing or creating social norms [Bicchieri and McNally, 2018, p.39]. This said, Bicchieri and McNally [2018, p.40-52] provide six examples where the success or failure in changing a script can explain why interventions succeeded or failed in creating or changing a social norm. Thus, we have preliminary evidence that the mechanisms described above can be used to explain how a social norm is created or activated in certain situations.

Finally, Cialdini et al. [1990] performed an experiment showing that a social norm of not littering can be activated by priming other cases of "proper behaviour" like recycling or voting. Their results support the claim that social norms are related in a semantic network, where focus on one social norm can activate a focus on other similar norms.

Given these considerations, I suggest that the mechanisms of categorisation and scripts provide a credible way for us to understand the emergence, activation, and change of social norms.

[10] See the references above.

The co-existence of two social norms in one situation

Using the above theories, we can now consider what will happen in situations where two social norms are present simultaneously. Recall that some situations can be categorised in more than one category. Depending on the categorisation, different scripts will be activated, and thus different norms can become salient for the person in the situation:

> In fact, the more ambiguous a situation appears to be, the greater the potential for conflicting interpretations, where each interpretation invokes a different norm. It is rather common for groups with conflicting interests to try to impose a reading of the situation that allows them to benefit from the application of a particular norm [Bicchieri, 2006, p.78].

The fact that people will often show a *self-serving bias* when choosing between two norms in a situation has been shown by Van Avermaet [1974] and Xiao and Bicchieri [2010].[11] Besides a self-serving bias, people's familiarity with different scripts, and the situational cues they focus on in the situation, are the most important factors for determining how a situation is categorised.

On this account, the framing of an ambiguous situation, along with the person's familiarity with different scripts, can have a strong effect on which norm people will end up following [Bicchieri, 2006, p.62]. This point is important in order to understand how microeconomics textbook can make their readers inclined to follow a social norm of (narrow) self-interest.

6.2.3 Reference networks

The final underpinning mechanism I will discuss is related to the use of *population* and *sufficiently large subset* in the definition of social norms. These concepts are important in order to argue how microeconomics teaching practices can stabilise a social norm of (narrow) self-interest. Here, I first present Bicchieri's different accounts of the concepts. Next, I use the current empirical literature to argue which account is best supported by data.

[11] See also Bicchieri [2006, pp.131-132] and Bicchieri [2016, pp.76-79].

Defining the population and each individual's threshold

In *The grammar of society*, Bicchieri [2006, pp.41,121-128] uses *population* to refer to a group of people. This group can range from a small group, where everyone knows each other, to a large group - like an entire country - where no one will know or observe everyone. Further, Bicchieri [2006, p.55] defines the *threshold* for when a person judges that sufficiently many follow the behavioural rule and expect the person to do the same (conditions 2(a) and 2(b)) as a percentage of the total population.

In *Norms in the wild* [2016], Bicchieri refines her account of population and how each individual's threshold is determined. Instead of being concerned with everyone in a group, Bicchieri [2016, p.53] introduces a *reference network* to account for the people whose actions and expectations an individual cares about when making a decision in a specific situation. The people in an individual's reference network can be physically present, but need not be [Bicchieri, 2016, pp.14,19]. An example of the latter is a famous singer who an individual might be influenced by despite never having met them. Further, the reference network can vary depending on the situation and the social norm in question. In some situations, the individual's family or co-workers might be the most important. In other situations, it might be some authorities on a specific topic, or the people that just happen to be where the individual is. The latter could be the case if there is an accident and the bystanders are considering what to do [Latane and Darley, 1968]. Finally, it may be the case that some people in the reference network are more important than others for the individual's expectations [Bicchieri, 2016, p.xiii].

If we want to identify a person's reference network, we can do this by asking them who they think will approve and disprove of a certain behaviour, along with who they would like to talk to about the behaviour [Fishbein and Ajzen, 2011, Bicchieri, 2016, ch.2, footnote 21].

A reference network can be represented by a graph, where *nodes* (illustrated by dots) represent people and where *edges* (illustrated by lines between the dots) represent a connection between two persons so that at least one

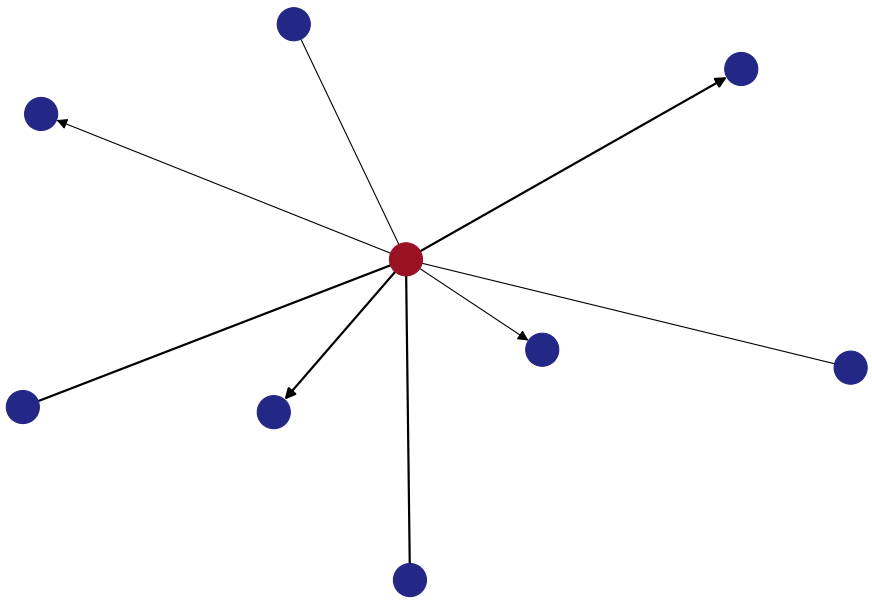


Figure 6.1: **Illustration of a reference network.** The red node represents the person whose reference network is illustrated for a given situation.

of the persons cares about the other person's expectations.[12] Figure 6.1 illustrates an example of a reference network in a given situation for one individual (the red node). Since a reference network represents the people whose expectations matter for an individual, this individual will be connected to everyone in the network. Note that four of the edges in the figure are directed. This indicates that the red node cares about the four nodes' expectations but not vice versa. Finally, we can illustrate how influenced the red node is by each of the blue nodes in its reference network by changing the *weights* of the edges. Here, I have illustrated the weight of an edge with its thickness where a thicker edge means that the nodes are more closely connected.

Given the introduction of a reference network, an individual will become a follower of a social norm if sufficiently many people in their reference network are believed to follow the norm and expect the individual to follow it (conditions 2(a) and 2(b)). What the threshold of *sufficiently many* is will

[12]Note that this use of "connected" does not imply that the two people know or have encountered each other.

vary between individuals and depend on each individual's allegiance to the norm (or norm sensitivity), risk sensitivity (how willing the individual is to take a risk in general), and risk perception of the specific situation the norm is associated with [Bicchieri, 2016, p.174].

Finally, Bicchieri [2016, p.xv,164,168,183] introduces a *social network* in order to discuss how a behavioural rule is adopted or changed for a group of people. A social network is a network that represents interactions between individuals [Serrat, 2017, p.40]. Interactions can, for example, be defined as sharing knowledge or communicating in some way (e.g. retweeting information), being linked through social media (e.g. being friends on Facebook), or being physically connected (e.g. by attending the same lecture at university). Bicchieri uses social networks to consider different communities where a social norm exists. Here the edges are, for example, defined by people who communicate with each other [Bicchieri, 2016, p.164]. Depending on the community, a social network can be a small, tightly knit, and isolated village where everyone interacts with everyone (so that the graph is fully connected) [Bicchieri, 2016, p.xiii] or it can be a larger community where people are not all connected, and where some people have more connections than others. Social networks make sense to consider, since people often care about the expectations of people they interact with. Thus, there will be an overlap between an individual's connections in a social network and that individual's reference network.[13] Note, however, that social networks differ from reference networks, first, because an individual may care about the expectations of a person who is not a part of their social network (like a religious authority on TV). Second, because an individual may be connected to a person in their social network who is not a part of their reference network. Thus, we should be careful not to confuse the two when talking about norm emergence and change.[14]

In order to avoid any confusion, I will introduce a *combined reference network* to describe a combination of a group of people's reference networks. In this

[13] We may interpret Bicchieri's 2006 account of population as a social network.

[14] Bicchieri [2016, pp.40,60,188,191] are examples where the two concepts are confused, or it is unclear from the text which network is discussed.

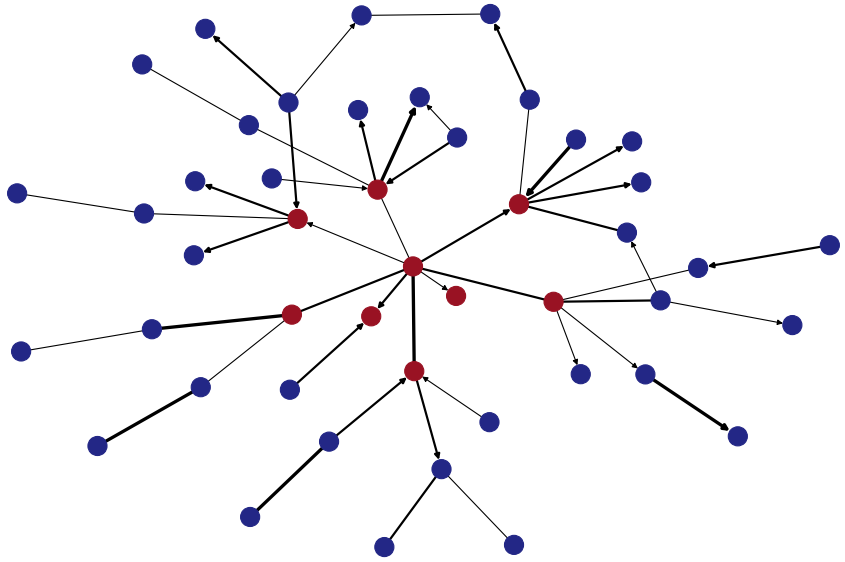


Figure 6.2: **Illustration of a combined reference network.** The red nodes indicate individuals who's reference networks are illustrated in the combined reference network.

network, the nodes represent people and the edges connect one person to another if the first person cares about the expectations of the other. Combined reference networks can be used to understand how individuals can affect each other with regard to a social norm due to their overlapping reference networks. An example of a combined reference network is presented in figure 6.2, where the thickness of the edges represent how much the nodes care about one another's expectations. Note that the network is not fully connected and that edges can be directed and non-directed.

Choosing the best model for social contagion

Whether we choose to use Bicchieri's definition of population from 2006 or 2016 will impact how social norms can spread in populations where everyone is not connected. Here, I will consider the evidence from different models for social contagion in order to determine which definition of population provides us with the most plausible account of how people can affect each

other with regard to social norms.

The literature on social contagion in networks is originally inspired by models of disease contagion. These early models assume that every time an individual is exposed to a disease, the individual will be infected by the disease with a certain probability. This is an example of a simple contagion model. When trying to model social behaviour, however, Granovetter [1978] suggested to use a *threshold model* instead. In a threshold model, an individual will only adopt a certain action, if sufficiently many other people have adopted that action. Threshold models are examples of *complex contagion* models and can vary in different ways. First, the threshold can be defined as an absolute number or as a fraction of the relevant scope for each node. Second, the scope for each node can be the entire network or the node's neighbourhood.

The model presented in Bicchieri [2006, ch.6] is an example of a threshold model where the threshold is defined as a fraction [Bicchieri, 2006, p.223] and where the relevant scope for each node is the entire population (or social network). However, several recent studies of online behaviour provide empirical evidence that complex contagion models - where the threshold is determined by the number of neighbours that adopt an action - fit well with real data on behavioural patterns [Centola, 2010, Kramer et al., 2014, Bond et al., 2012, Mønsted et al., 2017]. This model for contagion matches Bicchieri's [2016] account of reference networks and my suggestion to look at combined reference networks. Thus, I will use this framework to discuss the mechanisms that make an individual believe that sufficiently many people in a population will follow a behavioural rule and expect the individual to do the same (conditions 2(a) and 2(b)).[15]

[15]Note that since the threshold is different for each individual, it is possible to state each individual's threshold both as a the fraction and as an absolute number.

6.3 How microeconomics textbooks can make readers inclined to follow a social norm

In this section, I use the definitions and mechanisms discussed above to argue how microeconomics textbooks can make their readers inclined to follow a social norm of (narrow) self-interest in economic situations.

I do this in three steps. First, I argue that it is possible to have a social norm of (narrow) self-interest when using Bicchieri's definition of social norms. Second, I determine what counts as economic situations and argue that in all economic situations considered by the standard models, microeconomics textbooks either use a behavioural rule of narrow self-interest (if the opportunity for monetary gain is involved) or of self-interest (if the opportunity for monetary gain is not involved). Finally, I show how microeconomics textbooks can make their readers inclined to follow a social norm of (narrow) self-interest by making them satisfy Bicchieri's four conditions so that they will follow the norm unless they experience sufficient evidence that 2(a) and 2(b) are not satisfied in a situation. I end the final subsection by arguing how the fact that the same behavioural rules occur in most standard microeconomics textbook models can reinforce readers' beliefs that the norm is followed, and increase the types of situations in which the norm is assumed to apply.

6.3.1 Self-interest as a social norm

The aim of this subsection is to argue that it is possible to have a social norm of (narrow) self-interest given Bicchieri's definition of social norms. This is important since Bicchieri [2006, p.2] states that social norms are often opposed to people's narrow self-interest.

There are two ways one might argue that a behavioural rule of self-interest - and therefore narrow self-interest - cannot be a social norm in Bicchieri's framework. First, Bicchieri [2006, pp.11,26-27] states that social norms originate in situations that can be represented as mixed-motive games. They solve these situations by transforming them so that they can be represented

as coordination games, where cooperation (and not only self-interested behaviour) is a stable equilibrium. Thus, the entire purpose of social norms is to enforce prosocial behaviour and prevent people from following their own interests (utilities or monetary payoffs).

Second, since self-interested behaviour creates a stable equilibrium in situations that can be represented as mixed-motive games or coordination games, it can be argued that condition 2(b) is not needed to make people act that way. Further, Bicchieri [2006, pp.29,34] argues that descriptive norms (where conditions 1 to 2(a) are reasons for complying but condition 2(b) is not) are always dictated by self-interest. Thus, one might argue that a behavioural rule of self-interest will always be a *descriptive norm* rather than a *social norm*.

Answering the first argument, Bicchieri [2006, pp.2,25,29,34] does state that social norms often go against one's self-interest and especially one's narrow self-interest. However, she also argues that descriptive norms and conventions can become full social norms if - for example - breaking the coordination mechanism will cause severe negative externalities to other players [Bicchieri, 2006, pp.38-42,214]. Thus, the requirement that social norms enforce prosocial behaviour in situations that can be represented as mixed-motive games does not determine what will ultimately count as a social norm [Bicchieri, 2016, p.30].[16] Because of this, we cannot *a priori* exclude the possibility of a social norm of (narrow) self-interest.

The second argument can be answered in the same fashion: since it is possible for a descriptive norm to change into a social norm, the fact that self-interested behaviour initially best matches the definition of descriptive norms does not make it impossible for a social norm of (narrow) self-interest to exist.[17]

Finally, it is worth noting that Bicchieri [2006, p.9,189] refers to a social norm of self-interest.[18] Thus, it is possible for a social norm of (narrow) self-

[16] Bicchieri [2006, p.7] notes that many norms are not socially beneficial. For examples of harmful social norms, see Bicchieri [2016] and Bicchieri and McNally [2018].

[17] I discuss the empirical evidence for the claim that microeconomics promote a *social* norm of (narrow) self-interest in subsection 6.4.2.

[18] See also Miller and Ratner [1998] and Wuthnow [1991].

interest to exist on Bicchieri's account. Whether such a norm actually exists will depend on whether conditions 1 to 2(b) are satisfied for some people in some situations. In the next two subsections, I argue how the conditions can be satisfied for readers of microeconomics textbooks in economic situations.

6.3.2 Determining the behavioural rules used in economic situations

In this subsection, I use the analysis from section 5.3 to determine what will count as economic situations and which behavioural rules microeconomics textbooks use for each situation. This analysis will provide the first step for showing how microeconomics textbooks can make their readers inclined to follow a social norm of (narrow) self-interest in economic situations.

Looking at the microeconomic models discussed in section 5.3, we see that each model discusses what will happen in specific *types* of situations. In *consumer theory*, the model engages with types of situations where the agent is a consumer. This includes types of situations where the agent has to choose which goods to consume. Since the theory uses endowment (or income) to determine what the agent can consume, the model also covers types of situations where the agent has to consider how to spend their money, and whether to increase their income. In the *theory of the firm*, the model again engages with types of situations where the agent has to choose whether and how to increase their income.

For *choice under uncertainty*, the model engages with types of situations where the outcome of the agent's choice is uncertain. In *game theory*, the model engages with types of situations where the outcome of the agent's choice will depend on the choices of other agents. The standard games covered by microeconomics textbooks include both competitive situations and situations of coordination.

In *social choice theory* and *welfare economics*, the models consider types of situations where people are faced with a collective decision problem. Neither model, however, is concerned with how *individuals act* in collective decision problems. Thus, the textbooks do not use any individual behavioural rules

when discussing these situations. Since social norms relate to types of situations where individual behavioural rules can apply, I will not discuss situations involving collective decision problems any further.

Finally, in the theories of *competitive markets* and *general equilibrium*, the models engage with market and bargaining situations. Thus, microeconomics textbooks engage with several types of situations where an individual behavioural rule can apply. These types of situations can be summarised as situations involving:

- consumers or consumption,
- choices under uncertainty,
- coordination and competition with other agents, and
- bargaining or a market setting.

I will call these types of situations *economic situations*.

Next, we can use the analysis in section 5.3 to determine which behavioural rules microeconomics textbooks use for each economic situation. Here, I discuss each situation in turn.

Economic situations involving *consumers and consumption* are considered in the models of consumer theory, theory of the firm, competitive markets, and general equilibrium theory. Since competitive markets and general equilibrium theory use the model of consumers developed in consumer theory, it suffices to look at consumer theory and the theory of the firm in order to determine which behavioural rules microeconomics textbooks use in these types of situations. The variant of the rationality assumption present in consumer theory states that an agent faced with a choice between two consumption bundles will *always* choose the one they prefer to consume the most. Further, it states that the agent always will prefer more to less. Thus, microeconomics textbooks use a behavioural rule of self-interest to describe agent behaviour in consumer theory.[19] Since consumer theory assumes

[19]Indeed, it can be argued that it makes no sense to talk about other-regarding preferences in the model, since it is only concerned with one consumer and the goods they wish to consume.

that goods have a price, and an agent's utility will always increase with more goods, consumer theory also states that agents seek to maximise their income. Thus, in consumer or consumption situations that involve money, microeconomics textbooks use a behavioural rule of narrow self-interest to describe agent behaviour. This behavioural rule is also used for agents in the theory of the firm, since consumer theory provides the reason why firms are assumed to profit maximise.

Economic situations involving *choices under uncertainty* are discussed in theory of choice under uncertainty. Here, agents are assumed to maximise their own expected utility. This assumption need not imply a behavioural rule of self-interest. However, as discussed in section 5.3, this fact is not discussed in any of the microeconomics textbooks considered. Further, the textbooks only use examples of self-interested behaviour to explain the theory. Thus, microeconomics textbooks use a behavioural rule of self-interest in these situations, even though it is not required by the theory. Finally, in examples where the textbooks consider monetary payouts, they assume that agents act according to their narrow self-interest.[20]

Economic situations involving *coordination and competition with other agents* are considered in game theory, competitive markets, and general equilibrium theory. Here, I discuss the behavioural rules used by microeconomics textbooks in game theory, since the behavioural rules used in competitive markets and general equilibrium theory will be covered by economic situations involving bargaining and a market setting. In game theory, agents are again assumed to maximise their utilities. Even though it is not required by the theory, microeconomics textbooks only use examples where agents maximise their utilities by acting self-interestedly. Further, all textbooks use examples where the game payoffs are interpreted as money. Here, the textbooks use a behavioural rule of narrow self-interest to explain and predict the agents' actions.

Finally, economic situations involving *bargaining or a market setting* are discussed in the theory of competitive markets and general equilibrium

[20] Varian [2014, pp.217-222,126-233], Mas-Colell et al. [1995, pp.179-180,183-199,208-215], and Jehle and Reny [2011, pp.98-100,106-108,113-118,121-124].

theory. Since these theories draw on consumer theory and theory of the firm, the microeconomics textbooks use the behavioural rule of narrow self-interest in market situations (where money is involved) and self-interest in bargaining situations (where money is not involved). Thus, we see that for all economic situations, the microeconomics textbooks use a behavioural rule of self-interest if the opportunity for monetary gain is not present, and a behavioural rule of narrow self-interest if the opportunity for monetary gain is present.

6.3.3 Making readers inclined to follow a social norm

Microeconomics textbooks will make their readers inclined to follow a social norm if they make them satisfy conditions 1 to 2(b) in economic situations such that the readers will know that the social norm exists and follow it *unless* they are in a situation where they receive sufficient evidence that 2(a) and 2(b) are not satisfied. Here, I argue how reading microeconomics textbooks can make people satisfy each of the four conditions in economic situations so that they will be inclined to follow the social norm of (narrow) self-interest in economic situations. I end the subsection by suggesting how the likelihood of textbooks making their readers inclined to follow a norm can be increased by the use of models that consider different typical economic situations.

How conditions 1 to 2(b) can be satisfied

Condition 1: contingency Condition 1 requires that reading microeconomics textbooks will inform people that a behavioural rule of (narrow) self-interest exists and to which situations it applies. Looking at the previous subsection, we see that microeconomics textbooks use the behavioural rules of self-interest and narrow self-interest when discussing types of situations involving consumer and consumption, choice under uncertainty, coordination and competition with other agents, and bargaining or a market setting. Thus, readers of a microeconomics textbook will know that the behavioural rules of self-interest and narrow self-interest exist and that they can apply

to the economic situations for which they are used in the textbooks.[21] This information is sufficient to satisfy condition 1.

Notice that it is not given that a rule of (narrow) self-interest does or should apply in the above situations. This can for example be seen in the studies conducted by Yamagishi et al. [1998]. Here, they report that commitment formation - where buyers commit to buy from one seller and thus forgo the opportunity to look elsewhere for better prices - is present across cultures and in both low uncertainty and high uncertainty markets (though it is higher in high uncertainty markets). This shows that the relationship between one buyer and one seller can create a tie and a commitment such that the buyer prefers to stay with that seller even if it is possible to get a better price somewhere else. Thus, concerns for social ties and relationships can take priority over narrow self-interest in market situations.[22]

Condition 2(a): empirical expectations Condition 2(a) requires that microeconomics textbooks can make their readers believe that sufficiently many people will follow the behavioural rule in economic situations. There are at least two ways people can be prompted to believe this by reading a microeconomics textbook.

First, microeconomics textbooks can make their readers satisfy condition 2(a) because of the arguments and language used to defend the behavioural assumptions in these textbooks. Looking back at the discussion in section 5.4, I argued that contrary to Friedman's [1953] methodological instrumentalism, the textbooks defend the behavioural assumptions in the different variants of the rationality assumption as *descriptively plausible*. This can for example be seen by looking at the summarised arguments that include:

- It is reasonable, hardly objectionable, natural or almost tautological.
- Most economists do or assume it.
- It is typically the case, though there are a few counter examples.

[21] Note that I assume the readers are already familiar with the behavioural rules of self-interest and narrow self-interest so that microeconomics textbooks *promote* the rules in specific situations rather than *creating* them.

[22] See Bicchieri [2006, p.77] and Bicchieri [2016, p.114] for further examples.

- There is empirical evidence that it is the case.

Here, the textbooks use different rhetorical devises to claim that the behavioural assumptions are descriptively true. To give two examples, recall the text excerpts from section 5.3:

> Why would someone go to the considerable bother of creating a firm in the first place, and what guides such a person in the myriad decisions that must be made in the course of the firm's activities? Profit maximization is the most common answer an economist will give, and it is an eminently reasonable one. [...] These people are also consumers, and consumers get their satisfaction from the goods and services their income will buy [Jehle and Reny, 2011, p.125],

and:

> Although most people will usually prefer gambles that give better outcomes higher probability, as monotonicity requires, it need not always be so. For example, to a safari hunter [...] [Jehle and Reny, 2011, p.100].

In both cases, the readers of economics textbooks are informed that (most) people in economic situations will *actually* behave in accordance with the behavioural assumptions used in microeconomics. Further, the counter example provided by Jehle and Reny [2011, p.100] is a non-typical situation that is unlikely to have any relevance for the readers' own lives. Since the readers are informed that people *will* act in accordance with their (narrow) self-interest in economic situations, they are likely to form the belief that sufficiently many people actually do so. This in turn satisfies condition 2(a).

A more subtle way microeconomics textbooks can make their readers satisfy condition 2(a) is through the cognitive mechanisms of categorisation and scripts. As already noted, microeconomic models study *types* of situations. Specifically, the situations considered in each model are presented as being *typical* cases of each type that can provide us with insights:

> While the assumption of price-taking behaviour and the conditions leading to it are extreme, they provide a tractable model of the firm that is capable of yielding important insights. The

> competitive firm therefore merits our careful study [Jehle and Reny, 2011, p.145].

Thus, even when the textbooks state that the model assumptions may not apply to real situations, the readers are informed that the model can help us understand real situations of this type. This makes it possible for the readers to use the model as a *prototype* for understanding real situations.[23] Recalling the theory of categorisations, we use prototypes to categorise situations and to inform the scripts that we activate in these situations. If the readers come to believe that the model situations can be used to understand real situations directly, then they can adopt the model situations as prototypes that both help them categorise situations and provide them with scripts of what to expect in these situations. Since the textbook models assume that *everyone* (in the model) follows the behavioural rule of (narrow) self-interest, the scripts provided by these models will make the readers believe that this is also the case in real situations. In this way, microeconomics textbooks can provide their readers with prototype situations that will make them satisfy condition 2(a).

Condition 2(b): normative expectations Condition 2(b) requires that the readers believe that sufficiently many people *expect* them to follow the behavioural rule of (narrow) self-interest in economic situations. Here, the readers not only have to be convinced that everyone will behave according to the rule (as with a descriptive norm), but also that they are themselves *expected* to behave in this way. There are at least three ways that reading microeconomics textbooks can prompt this belief.

First, it is explicitly stated in game theory, competitive markets, and general equilibrium theory that everyone in the models behave rationally and that this is *common knowledge*. Thus, the agents in the models not only expect people to act rationally, they expect people to expect that they will all do so. Further, this assumption of *mutual expectations* about behaviour is necessary in order for the proposed solution concepts to work (such as

[23] See also Morgan [2012, pp.2-3,379,401-409].

dominant strategies in game theory, or the existence of clearing prices in general equilibrium theory). Thus, the model results depend on people in the models *expecting* everyone to act (narrowly) self-interested. This means that if the models are used to inform scripts of behaviour for the different situations they apply to, then the scripts will also include expectations that satisfy condition 2(b).

Second, as Bicchieri [2006, pp.43,96,ch.5 and 2016, pp.42-47] suggests, people may fall prey to the naturalistic fallacy, and infer from "everyone is doing this" to "everyone expects that everyone *ought* to do it". Recall the following excerpt from section 5.3:

> What would you think about a person who [violated transitivity]? This would certainly be taken as evidence of peculiar behavior.
>
> More importantly [...] [i]f we asked him to choose his most preferred bundle, he would have quite a problem, for whatever bundle he chose, there would always be one that was preferred to it [Varian, 2014, p.36].

Since the textbooks clearly state that people have reasons to follow the behavioural rule of (narrow) self-interest in the economic situations, the readers may internalise these reasons as more than just prudential, and come to believe that everyone agrees that everyone else ought to act in this way.[24]

Finally, as discussed in chapter 4, there is a long standing tradition in economic theory for claiming that it is *good* to follow one's (narrow) self-interest. This normative component of economics is also discussed by Morgan [2012, p.404] who writes:

> The theories, principles and laws of past economics embodied explanatory accounts at a rather general level, but they also - in their distinction between science and art - carried implicit normative suggestions about how the economy would or should behave, given the right governance. Models have inherited this positive (how the economic world is) and normative (how it should be) mix from earlier economics. But because models

[24] Note that even though this inference does not justify the belief satisfying condition 2(b), it might explain why the belief occurs.

> operate at a less general level than laws, they tend to embed the normative elements at a level closer to practical matters (however idealized the models themselves might appear to be). Indeed, it is this integration of the normative and positive aspects in models that prompts the way they are taken into the world and used directly as recipes to remake the world, and to change the behaviour of its people, as economists think it and they should function - that is, according to their models.

As discussed in section 5.4, it can also be seen directly in the textbooks, where arguments like

- We ought to act this way, or the theory can be used as a guide for action.

are used to defend the behavioural assumptions in the models as *normatively desirable*:

> A second advantage of the theorem is normative: expected utility may provide a valuable guide to action. People often find it hard to think systematically about risky alternatives. But if an individual believes that his choices should satisfy the axioms on which the theorem is based (notably, the independence axiom), then the theorem can be used as a guide in his decision process [Mas-Colell et al., 1995, p.178].

According to Bicchieri [2006, p.14] the belief that other people expect us to follow a behavioural rule is linked to the *normative expectations* that we *ought* to follow this rule. Since the textbooks argue that we *ought* to act in accordance with our (narrow) self-interest in economic situations, their readers can be persuaded to believe this claim. This, in turn, will provide them with the belief that we ought to act in this way, and that other people will therefore *expect* them to act in this way. Thus, condition 2(b) will be satisfied.

Condition 2: conditional preference Condition 2 requires that readers will *prefer* to conform to the behavioural rule *if* conditions 2(a) and 2(b) are satisfied. Whether this is the case will be an empirical matter, but it seems reasonable to assume that readers who believe that *everyone* follows a rule

of (narrow) self-interest in economic situations *and* believe that everyone expects them to do the same, will prefer to do so. This preference can for example be related to the wish to conform to people's expectations, to be a part of the group, or to avoid social dilemma situations where the reader is the only one who does not free-ride.

Summing up and further discussion

In summary, I have argued *how* microeconomics textbooks *can* make their readers satisfy condition 1, 2, 2(a) and 2(b) so that the readers will follow a social norm of (narrow) self-interest in economic situations. Notice that this will hold unless the readers get sufficient evidence that 2(a) and 2(b) are not satisfied in a situation. Thus, I have argued how microeconomics textbooks can make their readers inclined to follow a social norm of (narrow) self-interest in economic situations by affecting whether they satisfy conditions 1 to 2(b).

Using the theories of schema change and semantic networks, the likelihood of textbooks making their readers satisfy conditions 1 to 2(b) can further be increased in two ways due to the fact that microeconomic models present different prototypical situations that all include the same behavioural rules.

First, as noted in the bookkeeping model for schema change, the more something is repeated and stated as being right, the more likely we are to accept it as true and internalise it [Unkelbach et al., 2019]. Thus, the use of the same behavioural rule in several different models can make readers more prone to accept conditions 1 to 2(b) in each situation [Hawkins et al., 2001].

Second, using the same behavioural rule for different categories of situations can strengthen the ties between these categories in the readers' semantic network. Given the findings of Cialdini et al. [1990], this suggests that all categories related to economic situations will be activated if one of them is activated. This, in turn, will make it easier for the readers to access the economic categories and the script of (narrow) self-interest [Nosofsky et al., 2018].[25] The more easily a category is activated, the more likely the readers

[25]We can think of this process as analogous to training a muscle. The more the muscle or its neighbour muscles are being activated, the stronger it grows, and the more easy it is to

will be to interpret *ambiguous situations* as belonging to that category [Bicchieri, 2006, pp.216-217]. Further, Knez and Camerer [2000], Kay and Ross [2003], Mulford et al. [2008], and Peysakhovich and Rand [2016] observe a spill-over effect, where people apply a behavioural rule activated in one context to a new and different context. If the behavioural rule of (narrow) self-interest is more likely to be activated among readers of microeconomics textbooks, then this spill-over effect may make the readers apply the rule to situations that are not discussed by the textbooks. Thus, the fact that the behavioural rule of (narrow) self-interest is used in textbooks to study several types of situation, can increase the likelihood that their readers will follow the rule in ambiguous or new situations.

In the next section, I turn to the question of how microeconomics teaching practices can stabilise a social norm of (narrow) self-interest in economic situations among its students.

6.4 Stabilising a social norm via microeconomics teaching practices

In the last section, I argued how microeconomics textbooks can make their readers inclined to follow a social norm of (narrow) self-interest in economic situations. However, since social norms are contingent on 2(a) and 2(b) being satisfied, microeconomics textbooks readers may stop following the norm *if* they get sufficient evidence that conditions 2(a) and 2(b) are not satisfied in a given situation. Thus, whether a social norm will be stable over time, will depend on what people - who are inclined to follow the norm - will observe in real economic situations.

In this section, I argue how microeconomics teaching practices can stabilise a social norm of (narrow) self-interest by creating a reference network for the students in economic situations that will make them keep satisfying conditions 2(a) and 2(b). In subsection 6.4.1, I present my argument by using the mechanisms of reference networks, social networks, and combined

use.

reference networks discussed in subsection 6.2.3. In subsection 6.4.2, I argue how the claim that microeconomics textbooks and teaching practices can promote a social norm of (narrow) self-interest in economic situations is supported by the empirical findings discussed in chapter 3.

6.4.1 Using peers as evidence of norm following

Here, I argue how microeconomics teaching practices at universities can make students continue satisfying conditions 2(a) and 2(b) so that the norm will be stabilised for economic situations. I focus on university teaching practices in order to be consistent with the microeconomics textbooks discussed in sections 5.3 and 6.3. The argument has three parts. First, I discuss what a reference network for each student is likely to look like for economic situations. Second, I argue how microeconomics teaching practices can make a student continue satisfying conditions 2(a) and 2(b) in these situations. Finally, I suggest that combining the reference networks of the microeconomics students, we see how they can create a hub, that can spread the social norm to other parts of each student's social network.

Recall from subsection 6.2.3 that Bicchieri [2016, p.14] introduces the concept of a *reference network* to specify the relevant population that a person will care about when determining whether conditions 2(a) and 2(b) are satisfied in a given situation. When a student is in an economic situation - and has categorised it as such - I suggest that they will care about three groups of people: 1) people in the economic situation, 2) peers and teachers with whom they have discussed economic situations, and 3) friends and family who they want to please or conform with in general.

People in the economic situation are likely to be a part of the student's reference network because the student will be interacting with them.[26] Further, there are at least two reasons why we should assume that the student's peers and teachers in microeconomics will be a part of the reference network. First, the student has discussed how to act in economic situations

[26] Evidence that the people in a situation can influence one's choice can for example found in the literature on the bystander-effect [Fischer et al., 2011], group pressure [Asch, 1951], and social influence [Cialdini and Goldstein, 2004].

with this group. This has provided the student with direct information about their expected behaviour in these situations. Second, peers and especially teachers can be seen as experts with regard to economic situations. Thus, even if the student has discussed economic situations with their family, it is likely that they will give more weight to the teachers' assessment of how to act, since the teachers have authority with regard to economic situations.[27]

Since a student's reference network for economic situations is likely to include their fellow microeconomics students and teachers, it is relevant to consider whether microeconomics teaching practices include communication in a way that can make the student believe that their peers and teachers will follow a behavioural rule of (narrow) self-interest and expect others to do the same.

Economics teaching practices at universities generally consist of four activities. The first activity is reading selected parts of economics textbooks. The second activity is going to lectures, where the same selected parts are presented. The third part is class teaching, where a teacher will go through selected mathematical problems and proofs in order to provide mathematical support of the models and to train students in using them. Finally, students will be asked to solve problem sets either individually or in groups. For this final activity, one should expect students to discuss the problems and exchange advice regardless of whether the final assignment is to be handed in individually or in groups. Thus, all four parts of the teaching practices are aimed at making students understand, use, and accept the standard microeconomic models [Earle et al., 2016, pp.52-53].

Since understanding and using standard microeconomic models is the main aim of microeconomics teaching practices, a central conversation topic among students and with their teachers will be how to do this. The classes will focus on how to solve problems that apply the models to different (idealised) economic situations and derive the economic consequences of these applications. Further, due to the mathematical focus in microeconomics, there will not be sufficient time for teachers or students to discuss or question

[27] The influence on authority on people's decisions has famously been studied by Milgram [1963]. See also Blass [1999] and Burger [2009] for contemporary evidence.

the assumptions of the models.[28] Thus, the teaching practices encourage students to accept and apply the models from microeconomics textbooks in a non-critical way. Since the students are not exposed to alternative economic theories or to critiques of the neoclassical approach presented in textbooks, it is likely that each student will interpret the communication in classes and lectures as evidence that their fellow students and teachers will follow the behavioural rule of (narrow) self-interest in economic situations and will expect them to do the same. Depending on each student's threshold, the number of fellow students and teachers in their reference network may be sufficient to make them continue satisfying conditions 2(a) and 2(b).

If we assume that all students in a microeconomics class include their fellow students and teachers in their reference network, then we can create a combined reference network for the class, where the graph will be well connected between the students, and where the teachers of the class are likely to be in most students' reference network. An example of such a combined reference network is presented in figure 6.3. Here, the large red node in the centre represents an economics teacher. The red nodes surrounding it represent microeconomics students and the blue nodes represent people with no economic background that are still in the reference networks for some of the economics students in economic situations.

The students who have a low threshold may be convinced that a behavioural rule of (narrow) self-interest is followed in economic situations just by reading the microeconomics textbooks and listening to the lectures. These students will communicate their own empirical and normative expectations to their peers in the economics class (for example when trying to solve the problem sets together). This, in turn will convince a new set of students that more people will follow the rule of (narrow) self-interest in economic situations, thus bringing them closer to, or above, their threshold. As the course progresses, more and more students are likely to have their threshold met due to their interactions with their teachers and peers. Thus, a

[28] This claim is based on my own experience as an economics student as well as my experience as a microeconomics teacher. The assessment is also supported by the way the assumptions are (more or less) silently accepted in the microeconomics textbooks.

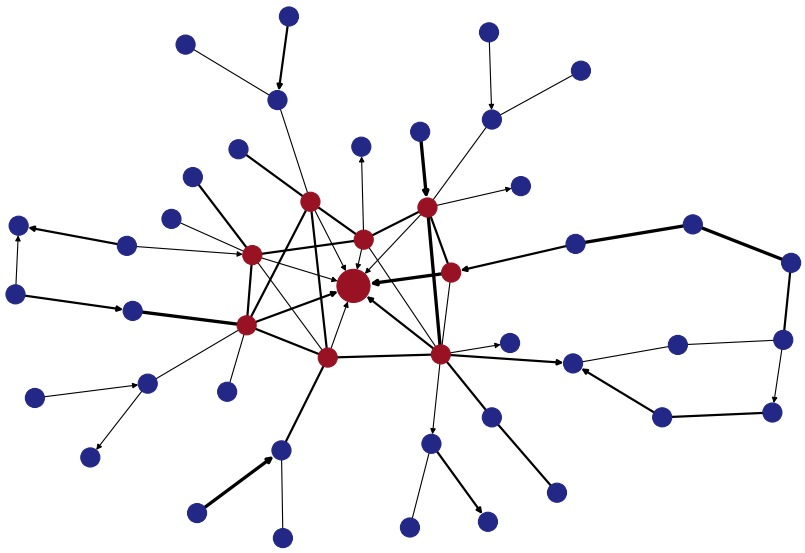


Figure 6.3: **Illustration of a simplified combined reference network for an economics class.** The large central red node node represents the teacher. The red nodes surrounding it represent economics students. The blue nodes represent non-economics students that still are in each student's reference network.

microeconomics class can create a *hub* in the combined reference network where the behavioural rule is confirmed to be followed and where this is common knowledge. From this hub, it is even possible that the students can affect their non-economic peers in their social network (who have economics students in their reference networks), thus making the social norm spread beyond the students themselves.

Summing up, economics teaching practices may create hubs of people among whom a social norm of (narrow) self-interest is stabilised for economic situations, so that the people in these hubs will always follow it. This can happen because the teaching practices promote a non-critical use and acceptance of the economic models, that can be interpreted as indicating that conditions 2(a) and 2(b) are satisfied in economic situations. Once a student observes that sufficiently many people in their reference network have accepted the norm, their threshold is met, and they will start following

the norm as well. Finally, from the basis of these hubs, the norm might spread to other parts of the students' social networks, assuming that the people with no background in economics have a sufficiently low threshold for norm following and that some of the students are a part of their reference networks.

6.4.2 Discussion: is the argument empirically plausible?

Combining sections 6.3 and 6.4, it is possible to give a full account of how the dissemination of microeconomics textbooks and teaching practices can make microeconomics self-fulfilling by promoting a social norm of (narrow) self-interest. I end this chapter by discussing how the experimental results presented in chapter 3 fit the above account.

First, the empirical results can be used to support the claim that microeconomics textbooks and teaching practices promote a social norm rather than a descriptive norm. According to Bicchieri [2006, p.150], social norms can overcome descriptive norms due to their normative component (condition 2(b)). The experiments discussed in chapter 3 show that dissemination of microeconomic theory and concepts can change people's behaviour in mixed-motive games [Ifcher and Zarghamee, 2018, Liberman et al., 2004, Kay et al., 2004, Molinsky et al., 2012]. From Engel and Rand [2014] and Fehr and Schurtenberger [2018] we also know that a social norm of cooperation already exists for non-economists in these games. If the norm of (narrow) self-interest promoted by economics had only been descriptive, it would have been overcome by the social norm of cooperation present in the situations. Thus, it has to be a social norm since it is able to crowd out another social norm [Bicchieri, 2006, pp.164-168].

Second, the argument that microeconomics becomes self-fulfilling by promoting a social norm can be used to explain two phenomena observed in the empirical literature. The first phenomenon is the finding that allowing participants to make promises before they play prisoner's dilemma will reduce the difference in behaviour between economists and non-economist [Frank et al., 1993, Hu and Liu, 2003]. This phenomenon can be explained

with the ability of one social norm to crowd out another social norm: in experiments where promise making is salient, the social norm of keeping promises is likely to crowd out the social norm of narrow self-interest. This pattern of social norms crowding out other social norms can also be seen for survey experiments and observational studies, where economists are prone to follow the social norm of (narrow) self-interest unless it is opposed by other social norms such as honesty [Laband and Beil, 1999, Yezer et al., 1996]. The second phenomenon that can be explained by applying Bicchieri's framework is the fact that economists and non-economists choose more similarly when more information about a situation is provided [Faravelli, 2007, Cipriani et al., 2009]. We can explain this by considering the theory of categorisation and scripts. If little information is provided, the person will apply the script they are most familiar with and which matches some of the cues in the situations. However, as the situation becomes less ambiguous, and several more situational cues are provided, the possible categorisations of the situation is reduced. Thus, the situation may no longer render itself to be categorised as an economic situation. This can for example explain the findings in Faravelli [2007, p.1401], where the difference in how economists and non-economists allocated resources (plants) between two individuals disappeared once it was explained that the difference in how well the individuals utilised the resources (by producing fruits) was caused by one of them being born with a physical disability.

In summary, the empirical literature presented in chapter 3 supports the argument that microeconomics becomes self-fulfilling by promoting a social norm. Further, the argument enables us to explain why there is a reduced difference between economists and non-economists in situations involving other social norms or when more information is provided about a situation. Thus, the empirical evidence discussed in chapter 3 fits the argument I have presented in this chapter.

6.5 Conclusion

In this chapter, I have used Bicchieri's definition of social norms to present a theoretical argument for how microeconomics textbooks and teaching practices can promote a social norm of (narrow) self-interest in economic situations.

First, I argued that it is possible to have social norm of (narrow) self-interest on Bicchieri's account of social norms. Next, I used the analysis of microeconomics textbooks presented in section 5.3 to determine what will count as an economic situation and which behavioural rules the textbooks use in these situations. Based on these findings, I argued how microeconomics textbooks can make their readers inclined to follow a social norm of (narrow) self-interest in economic situations by making them satisfy conditions 1 to 2(b).

Using the mechanisms of reference networks, I then argued how microeconomics teaching practices can stabilise a social norm of (narrow) self-interest among students so that they will keep following the norm. First, each student's peers and teachers are likely to be a prominent part of their reference network for economic situations. Second, teaching practices are constructed in a way that provides each student with evidence that their teachers and peers will follow a social norm of (narrow) self-interest in economic situations, and expect them to do the same. Thus, depending on each student's threshold, they can come to satisfy conditions 2(a) and 2(b) in all economic situations so that they will keep following the norm. Further, I argued that the microeconomics classes can create a hub of norm followers which may make it possible for the norm to spread to other parts of the students' social networks.

I ended the chapter by arguing that the above account of how microeconomics can become self-fulfilling is consistent with the empirical findings presented in chapter 3 and can explain why a few studies did not find a behavioural difference between economists and non-economists. Thus, the account fits the current empirical literature. In the next chapter, I present the results from three experiments conducted in order to test the plausibility of

the argument directly.

Chapter 7

Experimental evidence

7.1 Introduction

In chapter 6, I argued how microeconomics textbooks can make their readers inclined to follow a social norm of (narrow) self-interest by making them satisfy conditions 1, 2, 2(a), and 2(b) in economic situations. Thereby, readers will follow the norm unless they receive sufficient information to change their empirical and normative expectations (conditions 2(a) and 2(b)). Next, I argued how microeconomics teaching practices can stabilise the norm by making microeconomics students satisfy these expectations in economic situations such that they will keep following the norm.

In this chapter, I present the results from three experiments conducted to test the above arguments in order to argue *that* microeconomics textbooks and teaching practices can make people follow a social norm of (narrow) self-interest.[1] Combining the argument from chapter 6 with the empirical results in this chapter, we see that one way in which microeconomics can become self-fulfilling is by promoting a social norm of (narrow) self-interest via the dissemination of its textbooks and its teaching practices.

[1]The experiments were conducted in equal collaboration with Bjarke Mønsted and supervised by Sune Lehmann. See Buchter et al. [2020] which is also attached in the appendix to this thesis.

The chapter is divided into four sections. In section 7.2, I report the findings from the first experiment conducted to test whether the terminology used in microeconomics textbooks can change people's behaviour in a prisoner's dilemma game. In section 7.3, I present the results from a second experiment designed to test whether the behavioural changes are caused by changes in participants' empirical and normative expectations. Taken together, the results from the two experiments confirm that exposure to microeconomics textbook terminology can make participants inclined to follow a social norm of narrow self-interest. In section 7.4, I report the findings from a simulation experiment designed to study how the behaviour observed in the first experiment can have system wide consequences on the collective level if nodes modelled after this behaviour interact in a social network based on real student interactions. The findings from this experiment provide suggestive evidence that microeconomics teaching practices can stabilise a social norm of narrow self-interest among students. I end the chapter by discussing the experimental findings and addressing three potential worries.

7.2 Behavioural effects of microeconomic terminology

In this section, I provide experimental evidence from a controlled laboratory experiment to argue that participants' actions in an economic situation (the prisoner's dilemma) is conditional on their exposure to microeconomic terminology. This finding is important since it provides the first evidence needed to show that dissemination of microeconomic terminology can make people inclined to follow a social norm of narrow self-interest.[2] Further, the experimental results show that participants who have already been exposed to economic theory at university level defect more than participants with other educational backgrounds, independent of the terminology they are exposed to. Thus, in accordance with the empirical results discussed in

[2] I will return to this point in section 7.3.

chapter 3, these results confirm that there is a behavioural difference between economists and non-economists in a prisoner's dilemma situation.

The section is divided into two parts. In 7.2.1, I describe the experimental set-up and in 7.2.2 I report the results of the experiment.

7.2.1 Experimental set-up

The experiment consisted of each participant playing ten rounds of a prisoner's dilemma game (PD) on a computer. In order to avoid any potential effects of the name *prisoner's dilemma*, the game was referred to as *the game* throughout the experiment [Kay and Ross, 2003, Liberman et al., 2004, Kay et al., 2004, Belaus et al., 2018, Ellingsen et al., 2012, Dufwenberg et al., 2011].

Participants were informed that they would play against a new participant in each of the ten rounds. This meant that they could not expect previous rounds of the games to influence future rounds. Unbeknownst to the participants, they did not play against each other in the experiment. Instead, we programmed their computer to randomly choose cooperate or defect in each round. This was done in order to ensure that each participant's choices were independent of the choices of other participants such that we could distinguish between the effect of microeconomic terminology and the effect of playing against specific strategies. At the end of the experiment, participants were debriefed about the deception and the reason for it.[3]

The experiment was first conducted on *Amazon Mechanical Turk* (MTurk) and second in the *Behavioural Research Lab at London School of Economics* (BRL). The experimental design is the same for both studies and it was approved for the BRL study by the *LSE Research Ethics Committee*.[4]

[3]This kind of deception is generally discouraged within behavioural economics (Wilkinson and Klaes [2017, p.41] and Fiore [2009, pp.24-25]). However, recent voices in behavioural economics suggest that deception is acceptable if i) it is necessary for conducting the study, ii) it does not harm the participants, iii) participants are debriefed afterwards, and iv) the experimental setup has been approved by a research ethics committee [Bonetti, 1998, Cooper, 2014, Ortmann, 2019, p.35]. The experiment reported here satisfy all of the above requirements. Notice also that the methodological standards of other social science disciplines accept deception when the above conditions are satisfied [Gross and Fleming, 1982, Christensen, 1988, Krasnow et al., 2019].

[4]The remaining description of the experimental set-up along with the descriptions of the

Experimental design

The experiment had four parts. First, participants were presented with the title "A Choice Experiment" and read a general description of the experiment. Second, they were asked to read a description of the game and answer five control questions which ensured that they understood the rules of the game. Participants could not proceed to the game before all five control questions were answered correctly. Third, the participants played ten rounds of PD. After each round, participants were informed about the choice made by their opponent, their own pay-off from the game, and the choice and pay-off of another random participant.[5] The participants were then asked a follow-up question about the random participant in order to make them engage with this information. The participants could not continue to the next round of PD before they had answered the follow-up question correctly. Finally, after playing the ten rounds, participants were asked to state whether they had played this type of game before, and if they had guessed the hypothesis tested in the study. The experiment took less than 15 minutes to complete.

The PD played by the participants had the pay-off structure $T > R > P > S$ where $S = 0$, $P = 1$, $R = 2$, and T was selected uniformly at random from the interval $(2, 4)$. The two strategies were called *cooperate* and *defect*. The pay-off structure is depicted in figure 7.1. The entire experimental set-up can be seen at: http://ahura.herokuapp.com/.[6]

To test the effect of microeconomics textbooks terminology, participants were randomly allocated to one of three versions of the experiment. The control version used a neutral terminology and did not introduce a microeconomic concept. The second version used an individualist terminology and asked participants to read a text excerpt stating that in game theory the word *rational* is used to denote the strategy of defecting. The third version used a collectivist terminology and asked participants to read a text excerpt

the three terminologies are adapted from the SI in Buchter et al. [2020, pp.21-24].

[5]The information about the random participant was generated by the computer at random.

[6]To go to the experiment, enter a random sign in the field for an identification code, press "I agree - continue to study", and press "submit".

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	(2,2)	(0,T)
	Defect	(T,0)	(1,1)

Figure 7.1: **Structure of the PD played in the experiment.** The numbers indicate the points that players can win, where $T \in (2, 4)$ is chosen at random for each player. The value of T for each participant remains the same throughout the experiment.

stating that in game theory the word *optimal* is used to denote the strategy of cooperation. The two text excerpts were designed to be similar in their formulation, so that the only difference is whether they apply a positively laden word to the strategy of defection or to the strategy of cooperation. Both words relate to concepts used in microeconomics, though rational behaviour is more widely used than socially optimal outcomes.[7] Next, I provide a detailed account of the stimuli used in each of the three versions.

Neutral terminology

The first version of the experiment used a *neutral terminology* in order to provide a control version to which the other two terminologies can be compared. In this version, participants were greeted with the sentence "Welcome! You are about to take part in a study on how we make strategic choices." Further, "the other participant" was used to describe other players in the experiment. The versions did not introduce a microeconomic concept and the control questions pointed both to dominant strategies and to the benefit of cooperation without any normative wording:

1. If, in a given round, the other participant and you both play 'cooperate', how many points do you receive?

[7] See for example Mas-Colell et al. [1995, p.237]. Notice that the texts presented in the experiment provide a somewhat simplified version of game theory since rationality and social optimality is concerned with utilities rather than monetary gains. Thus, the concepts only apply to the experiment under the assumption that participants will increase their utility by increasing their wealth.

2. In a given round, you choose 'defect' and receive 1 point. Which strategy did the other participant choose?

3. If the other participant plays 'defect' in a given round, which strategy should you choose to ensure that you get the greatest possible number of points?

4. If the other participant plays 'cooperate' in a given round, which strategy should you choose to ensure that the two of you receive the greatest possible total number of points? Hint: for each strategy combination, sum the payoffs you and the other participant will receive.

5. Assume that in a given round you choose 'defect' and receive T. Which strategy did the other participant choose?

Finally the follow-up question after each round said:

- A random participant played [cooperate/defect] in the previous round, and received a payout of [T/2/1/0].
  Which strategy did that player's opponent choose?

This ensured that participants engaged with the information they received about the random participant's game.

Individualist terminology

The second version of the experiment used an *individualist terminology* to mirror the language of standard microeconomics textbooks. The terminology was introduced through four changes to the experiment.

First, the title of the experiment was accompanied by a small subtitle "A study on rationality" and participants were greeted with the sentence "Welcome! You are about to take part in a study on rationality." Further, "your opponent" was used to describe the other players. These situational cues were supposed to indicate to the participants that they were in a competitive situation.

Second, participants were asked to read a short text introducing the microeconomic concept of *rational* to describe the choice of defecting in the game. The text was:

> A concept of particular interest in this study is the notion of **rationality**. In game theory, we say that it is **rational** for a player to choose a strategy, if the strategy is guaranteed to result in a greater payoff to the player, regardless of which strategy their opponent plays. Conversely, we say that it is **irrational** for a player to choose a strategy that does not guarantee the highest possible payoff (regardless of what the other player chooses), if a strategy that does so is available. The following contains a few control questions to ensure that you understand these concepts and their relation to the game.

Third, control questions 3-5 were changed in order to ensure that participants understood the concept of *rationality* and knew how to apply it. The three control questions were:

3. If your opponent plays 'cooperate' in a given round, which strategy should you choose to ensure that you get the greatest possible number of points?

4. If your opponent plays 'defect' in a given round, which strategy should you choose to ensure that you get the greatest possible number of points?

5. Given your answers to the above, how would the 'defect' strategy be classified according to game theory?

Finally, the follow-up question after each round of the game was changed:

- A random participant played [cooperate/defect] in the previous round, and received a payout of [T/2/1/0].
  How does game theory categorize this strategy?

The participants could either answer *rational* or *irrational*. This change ensured that participants engaged with the information about the random participant and that they used the microeconomic terminology throughout the experiment.

Collectivist terminology

The third version used a *collectivist terminology*. The collectivist terminology was designed to mirror the individualist terminology by having a parallel sentence structure. The terminology was introduced through four changes to the experiment.

First, the title of the experiment was accompanied by a small subtitle "A study on cooperation" and participants were greeted with the sentence "Welcome! You are about to take part in a study on cooperation." Further, "your co-player" was used to describe the other participants in the experiment. These changes were intended to provide the participants with a situational cue that they were in a cooperative situation.

Second, participants were asked to read a short text introducing the concept of *optimal* to describe the choice of cooperating in the game:

> A concept of particular interest in this study is the notion of **social optimality**. In game theory, we say that an outcome is socially optimal if it results in the largest overall payoff and if no one can be made better of without making someone else worse off. We call a strategy that can lead to a socially optimal outcome **optimal**. Conversely, we call a strategy **suboptimal** if it cannot lead to a socially optimal outcome. The following contains a few control questions to ensure that you understand these concepts and their relation to the game.

Third, control questions 3-5 were changed to ensure that participants understood the concept of optimality and knew how to apply it:

3. If your co-player plays 'cooperate' in a given round, which strategy should you choose to ensure that the two of you receive the greatest possible total number of points (i.e. which choice maximizes the sum of the points that you and your co-player receive)?

4. If your co-player plays 'defect' in a given round, which strategy should you choose to ensure that the two of you receive the greatest possible total number of points? (i.e. which choice maximizes the sum of the points that you and your co-player receive?)

5. Given the above how may we classify the role of the 'cooperate' strategy in increasing overall wealth?

Finally, the follow-up question was changed to:

- A random participant played [cooperate/defect] in the previous round, and received a payout of [T/2/1/0].
  Which of the following best describes this strategy choice?

The participants could either answer *optimal* or *suboptimal*.

The experiment was conducted with two groups of people: one on MTurk and one in BRL. Before turning to the results of the experiment, I briefly provide information on the two groups.

Data source 1: MTurk

There were 344 participants in the study. In order to secure data quality, only highly rated MTurk workers were allowed to participate [Lee et al., 2018, Hauser and Schwarz, 2016, Peer et al., 2014]. No demographic restrictions on workers were made [Lee et al., 2018]. The study ran in November 2018 and all participants had completed the study within one day of launching it. Participants were paid 2 USD as a base rate for participating. In addition, they earned 10% of the points they collected during the experiment in USD. The average earnings were 3.7 USD with a minimum earning of 2.0 USD and a maximum earning of 5.6 USD. The median earning was 3.6 USD. Of the 344 participants, 72 reported that they had played a similar game previously.

Data source 2: BRL

There were 466 participants in the study. Of these, 402 stated that they were students (either full time of part time), 311 identified as female, 154 as male, and one identified as other. The age of the participants was between 18 and 72, with an average age of 24.3 and a mean age of 22.

Looking at participants' educational background, 77 participants associated with one of three degrees that include at least two years with economics

courses at LSE. The degrees were Economics, Finance, and Accounting.[8] For simplicity, I will refer to this group of participants as *economists*. The label is based on the hypothesis that this group has been exposed to more microeconomics than other participants.[9] In the group of economists, 51 identified as females and 26 identified as males. The researchers and assistants did not know the educational background or any other non-observable demographic background information of the participants during the study.

In the group of economists, 22 reported having played a similar game earlier. This number is 55 for participants not belonging to the group of economists. For simplicity, I will call this group *non-economists*.

The experiment in BRL was conducted in the first week of December 2018. Participants completed the study in sessions of up to 20 people at a time. Participants were allowed to join a session up to five minutes after the start. Thus, participants could not join the study after people started playing the prisoner's dilemma game.

Participants earned a base pay of 5 GBP for participating in the study. In addition, they earned 10% of the points they collected during the game in GBP. The average earnings were 6.7 GBP with a minimum earning of 5.6 GBP and a maximum earning of 8.2 GBP. The median earning was 6.7.

7.2.2 Results

Before analysing the behavioural effects of the terminologies, we first compared the two data sources. Demographics studies of MTurk workers show that the majority of workers live in USA and provide a general sample of the population (but with a slightly lower average household income) [Difallah et al., 2018, Hara et al., 2019]. Looking at the current demographics in the

[8] Looking at the LSE course description for undergraduates, only degrees that have "economics", "finance", or "accounting" in their title have two or more years with economics courses. Participants were not able to report degrees that combine fields of study, and thus had to select the one programme they identify with the most.

[9] The measure is limited since it is possible for people on master's level to have had economics in their undergraduate degree and changed to another degree at master's level. This limitation, however, will only make our results stronger, if there is a significant difference between the two groups.

US, 2.3% of the population has a bachelor's degree in the social sciences (of which economics is a part).[10] In contrast, 17% of the participants in the BRL sample study or have studied economics. Because of the skewed number of economists in the two sample and since the results provided in chapter 3 show that economists defect more than non-economists in PD, we hypothesised that participants from BRL would defect more than participants from MTurk.

Analysing the two data sources, there is a significant difference between the MTurk data source and the BRL data source with BRL participants defecting more across the three versions of the experiment: $p = 0.048$, $t = -1.66$ (linear regression, one-tailed). This difference, however, disappears when the group of economists are excluded from the BRL data set: $p = 0.25$, $t = -0.66$ (linear regression, one-tailed).[11] Thus, we can get an understanding of how the general population (US/UK) will respond to the three terminologies by looking at the combined data set of MTurk and BRL where the 77 economists in the BRL data source are excluded. This combined data set consists of 733 participants.

Based on the above finding, data from the experiment will be analysed in two steps. First, the behavioural effect of the three versions of the experiment will be analysed using the combined data set excluding economists. This analysis shows that participants' choices are conditional on their exposure to the different terminologies. Second, the behavioural difference between economists and non-economists in the BRL data set is analysed. This analysis shows that economists defect more than non-economists, that the difference is caused by participants' educational background, and that economists and non-economists are equally influenced by the three terminologies.

[10] This calculation is based on the 2018 US work force population being 163.1M (https://fred.stlouisfed.org/series/CLF16OV) and the 2018 workforce of social scientists being 3.82M (https://datausa.io/profile/cip/economics).

[11] The fact that there is no significant difference between the data collected from MTurk and BRL is in alignment with several studies conducted on MTurk data quality [Kees et al., 2017, Hauser and Schwarz, 2016, Harms and DeSimone, 2015, Klein et al., 2014, Peer et al., 2014, Berinsky et al., 2012, Rand, 2012, Buhrmester et al., 2011].

Behavioural effects of different terminologies

In order to analyse whether participants' choices are conditional on which terminology they are exposed to, we used the combined MTurk and BRL data set of 733 participants.

Overall, the number of times participants chose to defect were significantly affected by which terminology they were exposed to: $p = 6.8 \cdot 10^{-9}$, $H = 38$ (Kruskal-Wallis test). Looking at the difference in defection between participants in each version of the experiment, we see that participants exposed to the individualist terminology defected significantly more than participants exposed to neutral terminology: $p = 8.2 \cdot 10^{-6}$, $t = 4.34$ (one-tailed Conover-Iman post-hoc analysis [Conover and Iman, 1979]). Further, participants exposed to collectivist terminology defected less than participants exposed to neutral terminology: $p = 0.03$, $t = 1.89$ (one-tailed Conover-Iman post-hoc analysis).[12] The average number of times participants exposed to each terminology defected are shown in figure 7.2.a [Buchter et al., 2020, Fig. 1, p.4].[13]

Figure 7.2.b shows the fraction of participants who defected in the three versions of the experiment for each of the ten rounds. As we can see, participants in all versions became more likely to defect the more rounds they played. However, we also see that the difference between the versions are

[12]Looking at **Mturk participants** only, the terminologies significantly affected how many times people defected: $p = 3 \cdot 10^{-5}$ with $H = 21$ (Kruskal-Wallis test). Participants defected significantly more in the individualist version compared to the neutral version: $p = 1.3 \cdot 10^{-3}$, $t = 3.04$ (one-tailed Conover-Iman test) and significantly less in the collectivist version compared to the neutral version: $p = 0.039$ with $t = 1.77$. For **BRL participants** (including economists) the terminologies significantly affected how many times participants defected: $p = 6 \cdot 10^{-7}$, $H = 29$ (Kruskal-Wallis test). Participants in the individualist version defected significantly more than participants in the collectivist version $p = 4 \cdot 10^{-5}$, $t = 3.98$ (one-tailed Conover-Iman test). However, there was no significant difference in the amount of defection between the collectivist and neutral version of the experiment ($p = 0.11$, $t = 1.22$).

[13]Notice that the effects of the terminologies are more significant when economists are included in the data set: the overall effect of terminologies on defection is significant with $p = 0.02$, $t = 2.06$ (Kruskal-Wallis test). Participants defect more in the individualist version compared to the neutral version ($p = 3.5 \cdot 10^{-12}$, $t = 6.96$) and defect less in the collectivist version compared to the neutral version ($p = 3.7 \cdot 10^{-7}$, $t = 4.99$ one-tailed Conover-Iman post hoc test). The same tests for BRL economists alone give us the results $p = 0.063$, $t = 1.55$; $p = 3.2 \cdot 10^{-5}$, $t = 4.24$; and $p = 9 \cdot 10^{-3}$, $t = 2.43$ respectively.

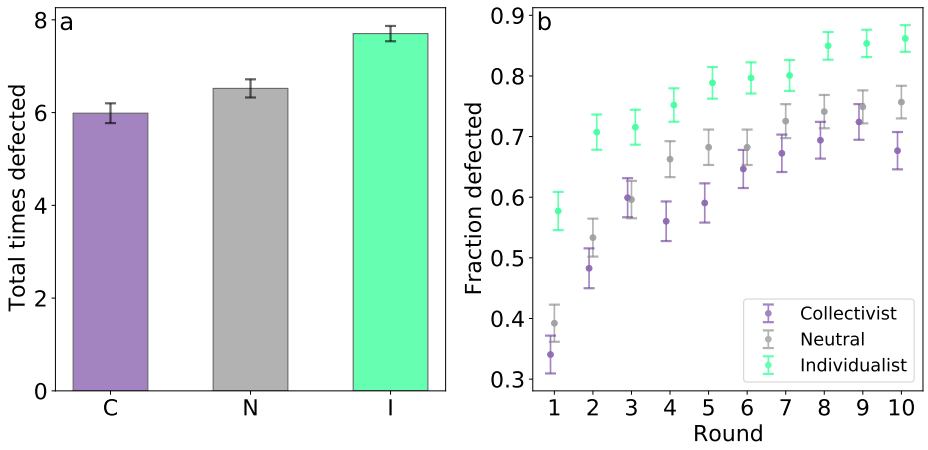


Figure 7.2: **Effect of terminology on behaviour. a** The average number of times participants in each of the three versions of the experiment (collectivist (C), neutral (N), and individualist (I)) defected. **b** The fraction of participants who defected in each round for each terminology.

more or less maintained in spite of all participants playing against strategies generated at random.

Finally, we tested whether an increase of T in the payout matrix increased the number of times participants defected. We see a significant effect of increasing T in the first round: $p = 8.83 \cdot 10^{-5}$, $z = 3.75$ (logistic regression, one-tailed z-test) and for all ten rounds: $p = 0.024$, $t = 1.98$ (linear regression, one-tailed t-test). Looking at the effect of previous game experience, we see a small effect in the first round of the game: $p = 0.034$ (logistic regression, one-tailed z-test) but no effect when all ten rounds are considered $p = 0.12$ (linear regression, one-tailed t-test).[14]

The results show that the number of times participants defect is conditional on the terminology they are exposed to. Further, we see that the difference in choices persist across the ten rounds participants played. Since all participants played against strategies chosen at random and since their choices are independent of each other, this confirms that the behavioural effects are

[14]The effect of previous game experience is significant for MTurk participants ($p = 1.3 \cdot 10^{-3}$, $t = -3.034$ one-tailed t-test) but not for BRL participants (including economists): $p = 0.705$, $t = 0.378$ (one-tailed t-test).

caused exclusively by their exposure to terminologies.

I end this section by analysing the behavioural difference between economists and non-economists observed in the BRL data set.

Economists

Analysing the BRL data set only, we see that economists defect more than non-economists across the three versions of the experiment: $p = .0007$, $t = -3.22$ (linear regression, one-tailed t-test).[15] Looking at each terminology separately, there is a significant difference in the median number of times each group defected for the individualist terminology and for the neutral terminology with economists defecting more than non-economists: $p = 0.000243$ and $p = 0.0284$ (Mann-Whitney U-test). There is no significant difference for the collectivist terminology: $p = 0.105$ (Mann-Whitney U-test). The average number of times economists and non-economists defected are shown in figure 7.3 for each of the three versions of the experiment.[16]

Considering the *effects* of the terminologies on each group we see that both groups are significantly influenced by the terminologies they were exposed to: $p = 4.9 \cdot 10^{-4}$ for economists and $p = 1.4 \cdot 10^{-4}$ for non-economists (Kruskal-Wallis test). Further, there is no significant difference between how much the groups are affected by the terminologies when comparing participants in the collectivist version and the neutral version ($p = 0.84$) nor when comparing participants in the individualist version and the neutral version ($p = 0.85$, linear regression looking at all ten rounds). Thus, the results show that the two groups are *equally influenced* by the different terminologies, but that economists in general defect more than non-economists.

Finally, figure 7.3 indicates that the extent to which non-economists defect in the individualist version corresponds to how much economists defect in the neutral version. Looking at a linear regression model, we see that economists defect 1.12 times more than non-economists during the ten rounds, regardless

[15] Controlling for education and terminologies, a linear regression on the number of times participants defected during the ten rounds show a significant effect of age, with younger people defecting more $p = 0.01$. There was no effect of participants' student status ($p = 0.98$) nor of participants' gender ($p = 0.29$).

[16] For the difference across the ten rounds see Buchter et al. [2020, Fig.1, p.4].

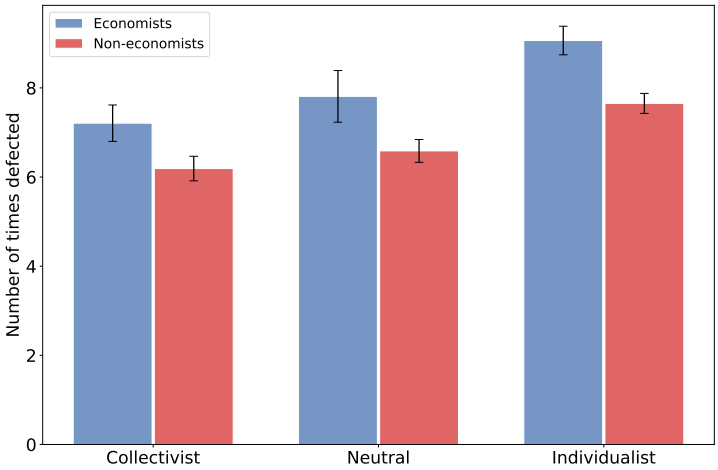


Figure 7.3: **Difference in defection between economists and non-economists.** The average amount of defection across the ten rounds for economists and non-economists in the BRL data set for each of the three versions of the experiment.

of the terminology. Further, we see that the individualist terminology makes all participants defect 1.15 times more.[17] Thus, the behavioural effect of the individualist terminology corresponds to the general behavioural difference between economists and non-economists.

In order to test whether the increased defection rate among economists can be explained by other factors than their education, we first tested whether the group of economists were influenced by T. Contrary to the analysis of the larger data set, there is no significant effect of T for economists: $p = 0.911$ (logistic regression on the first round) and $p = 0.422$ (linear regression on all rounds). Testing for the amount of errors made in the follow-up questions, the time spent on follow-up questions, and the time spent on choosing a strategy in each round of the game, we found no interesting differences between economists and non-economists in the three versions of the experiment.[18]

[17] The model considers the effect of the collectivist terminology, the individualist terminology, whether the participant belongs to the group of economists, T, age, previous experience, and student status.

[18] **A.** There is no significant difference between the groups on how much their defection rates are correlated with number of mistakes in the follow-up questions for each of the three terminologies: $p = 0.671$, $p = 0.149$, and $p = 0.461$ (using Pearson correlations). **B.** Considering the difference - using Pearson correlations - between the two groups in time

Thus, the experimental results suggest that the determining factor for the behavioural difference between economists and non-economists is their education.

7.3 Do terminologies promote social norms?

In *Norms in the Wild*, Bicchieri [2016, ch.2] argues that two independent measures are needed in order to determine whether a social norm exists in a situation and has causal power, such that it will be followed when people's empirical and normative expectations are met. The first measure should determine what people's empirical and normative expectations are for the situation. The second measure should test whether people's behaviour changes conditional on their empirical and normative expectations being met [Bicchieri, 2016, p.51]. The measures should be independent since making participants formulate their expectations may influence how they choose to behave. Combining the two measure will show whether conditions 2(a), 2(b), and 2 are met for the people in the situation.

The terminologies in the first experiment were designed to provide participants with empirical and normative cues about the situation they are in. The empirical cues were constructed via the small changes in framing of the experiment (co-player/participant/opponent). The normative cues were provided by the use of the microeconomic concepts of *rationality* and *social optimality*. Thus, the first experiment tested whether the terminologies - providing empirical and normative cues about the situation - could affect people's behaviour.[19] [20] The exposure to the individualist terminology

spent before answering the follow-up questions, there is no significant difference in the collectivist and individualist versions of the experiment ($p = 0.358$ and $p = 0.11$), but there is significant difference for the neutral narrative ($p = 0.0443$). **C.** Finally, when considering the Pearson correlations between how much an individual defect and the median time spent on making a decision during the ten rounds of the game, we found no significant difference between the two groups: $p = 0.279$, $p = 0.988$, and $p = 0.267$.

[19] In accordance with Bicchieri [2016, p.59,67] the experiment was performed as a controlled laboratory experiment.

[20] We did not test the effect of the empirical cues and normative cues independently since this has already been done in previous experiments (see section 3.5). I will discuss the use of both empirical and normative cues in section 7.5.

tested whether participants could be encouraged to follow a behavioural rule of defection. The exposure to the collectivist terminology tested whether participants could be encouraged to follow a behavioural rule of cooperation. The results from the experiment show that participants' behaviour *is* conditional on which terminology they are exposed to. Thus, *if* the terminologies manipulate participants' empirical and normative expectations, *then* the first experiment shows that two social norms exist and have causal power in the prisoner's dilemma situation and that condition 2 is satisfied both for a social norm of defection and for a social norm of cooperation.[21]

However, in order to conclude that the behavioural changes observed in the individualist and collectivist terminologies are caused by the presence of two social norms, we need an independent test of whether the behavioural changes are caused by changes in participants' empirical and normative expectations. This is done in a second experiment conducted in the spring of 2019, where participants were asked about their empirical and normative expectations before playing each round of PD.

In this section, I report the results from the second experiment. In subsection 7.3.1, I describe the experimental design and data source.[22] In subsection 7.3.2, I report the results of the experiment and discuss how they fit the hypothesis that the behavioural changes observed in the first experiment are caused by the promotions of a social norm of defection and a social norm of cooperation, respectively.

7.3.1 Second experimental set-up

In order to make sure that the situation in the second experiment is comparable to the situation in the first experiment, we used the same experimental design as reported in 7.2.1. However, we made one change to the experiment. In each round of the PD, participants were asked two additional questions, before they indicated which strategy they wanted to play. The first question

[21] The co-existence of two social norms in a situation like the prisoner's dilemma has been discussed in subsections 6.2.2 and 6.4.2.

[22] The description of the experimental set-up is adapted from the SI in Buchter et al. [2020, pp.24-25].

was designed to ask about the participants' empirical expectations (condition 2(a)), while the second question was designed to ask about participants' normative expectations (condition 2(b)). The questions were:

- Which strategy do you think the other participant will choose?
- We also ask the other participant which strategy they think **you** will choose. What do you think the other participant answers?

The participants could answer *cooperate* or *defect* to each question. For participants in the collectivist version of the experiment, "the other participant" was changed to "your co-player". For participants in the individualist version of the experiment, it was changed to "your opponent".

Data source

The second experiment was conducted on MTurk, using only highly rated MTurk workers. As before, no demographic restrictions on workers were made. The study ran in March 2019, and it was completed within one day of launching it. As in the first MTurk experiment, participants earned a base rate of 2 USD for participating and an additional 10% of the points they collected in the game in USD. The average earnings were 2.84 USD with a minimum earning of 2.14 USD, and maximum earning of 3.62 USD. The median earning was 2.80 USD. Of the 200 people participating, 50 reported to have played a similar game previously.

7.3.2 Results

Figure 7.4 shows the behavioural differences between the three versions of the second experiment. In figure 7.4.a we see that participants exposed to the collectivist terminology defect significantly less than participants exposed to the neutral terminology ($p = 6.5 \cdot 10^{-4}$, $t = 3.26$, one-tailed Conover-Iman posthoc), but that the difference between the neutral and individualist terminology is not significant ($p = 0.102$, $t = 1.28$).[23] The difference in

[23] The difference between all three terminologies is significant with $p = 1.2 \cdot 10^{-4}$, $H = 18.08$ (Kruskal-Wallis test).

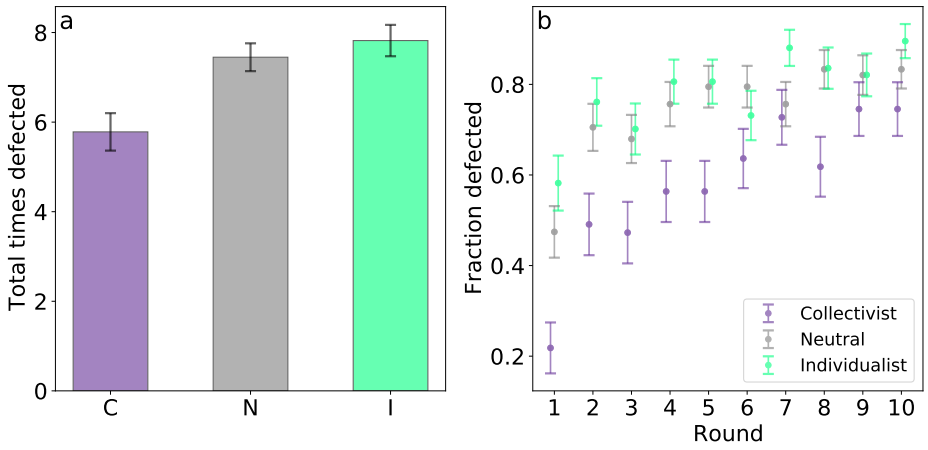


Figure 7.4: **The behavioural differences between the three versions of the second experiment.** **a** The average number of times participants in each of the three versions of the experiment defected throughout the ten rounds. **b** The fraction of participants in each of the three versions who defected for each of the ten rounds in the experiment.

significance between the neutral and individualist terminology may be due to the smaller number of participants, or to the fact that asking participants to formulate their empirical and normative expectations can influence their choices. Looking at figure 7.4.b, we see that the fraction of participants who defected increases during the ten rounds, but that participants in the collectivist version generally kept cooperating more than participants in the other versions of the experiment.

Comparing the results from the second MTurk experiment to the results from the first MTurk experiment, we see that there is no significant difference between the two data sets: $p = 0.13$, $t = -1.50$, two-tailed linear regression.[24] Thus, we find it safe to use this experiment to explore why the behavioural differences we observed in the first experiment occur.

In order to consider the relation between social expectations and behaviour across all three terminologies, we first examined whether social expectations (satisfying both condition 2(a) and condition 2(b)) with regard to either

[24]The regression had the number of times participants cooperated as the target variable and the terminologies and the MTurk data sets as their input variables.

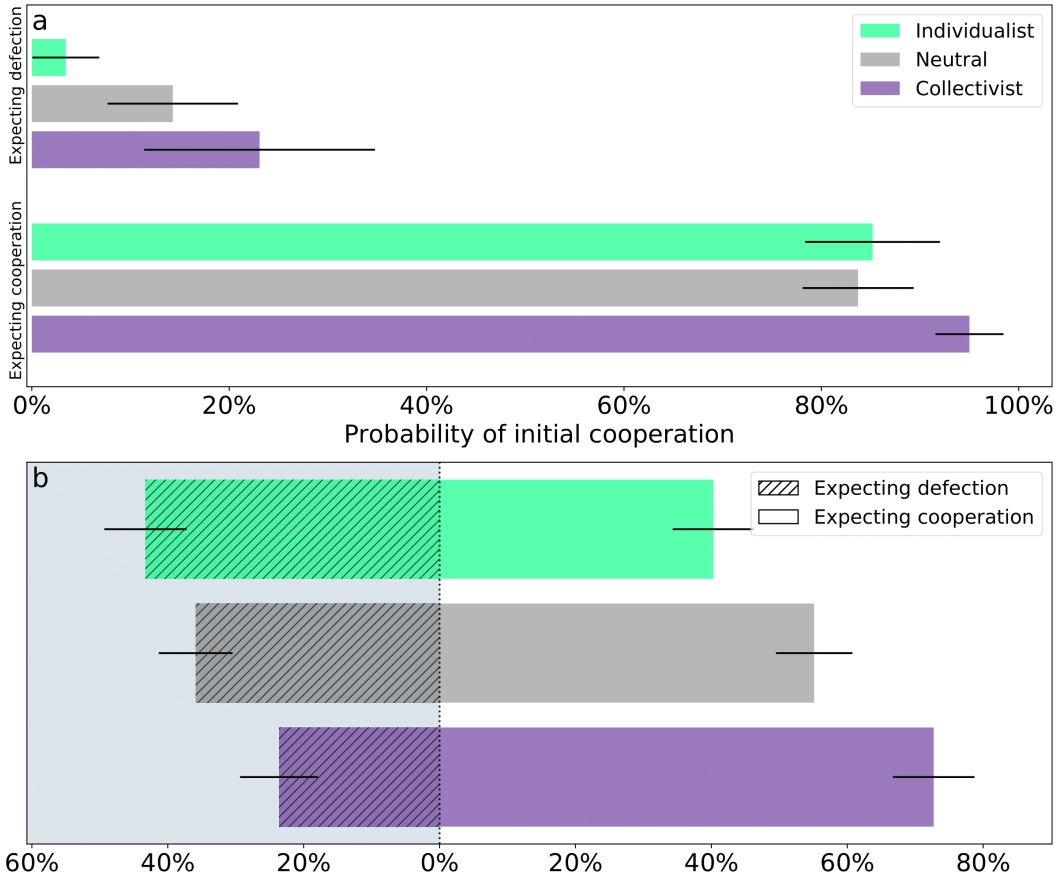


Figure 7.5: **The interplay between expectations and terminology exposure in the first round. a** The percentage of participants choosing to cooperate in the first round, grouped by their expectations and terminology exposure. "Expecting cooperation" means that conditions 2(a) and 2(b) are satisfied for cooperation. "Expecting defection" means that conditions 2(a) and 2(b) are satisfied for defection. The expectations for each version of the experiment need not sum to 100% as some participants gave different answers to the two questions. **b** The probability of participants expecting defection or cooperation, respectively, in the first round of the experiment for each of the three terminologies.

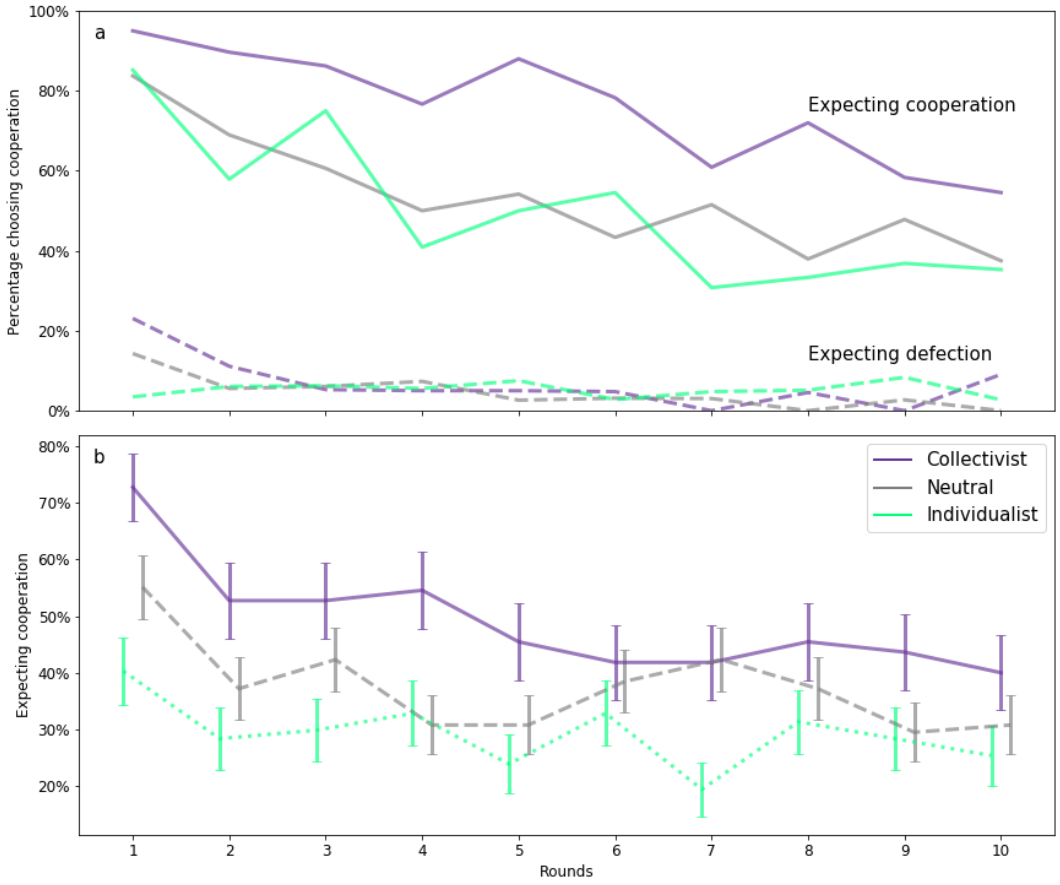


Figure 7.6: **The interplay between expectations and terminology exposure during all ten rounds.** **a** The percentage of participants choosing to cooperate in each round, grouped by their expectations and terminology exposure. **b** The probability of participants expecting cooperation for participants in each of the three version of the experiment. The error bars are given by $\sigma_f = \sqrt{(f\,(1-f))/n}$.

cooperate or defect are a good predictor of which strategy participants chose. Looking at the first round only - where noise from the game has not yet been introduced - figure 7.5.a shows that participants who expected cooperation (satisfied 2(a) and 2(b) for cooperation) were likely to cooperate, while participants who expected defection (satisfied 2(a) and 2(b) for defection) were likely to defect [Buchter et al., 2020, Fig.2, p.6]. Across the three versions of the experiment, 88.2% of participants who expected cooperation in the first round chose to cooperate, and 88.8% of the participants who expected defection in the first round chose to defect. The difference is significant with $p = 1.2 \cdot 10^{-24}$, $z = 10.18$ (proportions z-test). Figure 7.6.a, depicts the same relation as figure 7.5.a, but for all ten rounds. These findings indicate first, that participants' social expectations are a good predictor of how they choose to act. Second, the results support the hypothesis that two social norms exist in the experimental situation and that participants will follow the norm which meets their social expectations.

Next, we tested how the terminologies affected participants' social expectations. First, we examined whether the terminologies increased participants' likelihood of acting in accordance with their expectations such that participants in the individualist (collectivist) version who expected defection (cooperation) would be more likely to defect (cooperate). As can be seen in figure 7.5.a, the terminologies do not affect how likely participants who expect defection are to defect in the first round (comparing the individualist terminology to the neutral: $p = 0.84$, $z = -0.99$, one-tailed proportional z-test). Similarly, the terminologies do not affect how likely participants who expect cooperation are to cooperate in the first round (comparing the collectivist and neutral terminology: $p = 0.14$, $z = 1.09$, one-tailed proportional z-test). Looking at all ten rounds, figure 7.6.a shows that there is a small effect of the collectivist terminology on how much participants with cooperation expectations cooperate. However, there is no effect of the terminologies with regard to the defection rate of people who expected defection. Thus, the terminologies do not have a large effect on participants' propensity to act in accordance with their expectations.

Second, we tested whether the different terminologies influenced which

social expectations participants had. Here, the number of participants expecting defection and the number of participants expecting cooperation were significantly influenced by the terminologies across the ten rounds: $p = 0.044$, $H = 6.3$ and $p = 0.003$, $H = 12$, Kruskal-Wallis test. The influence of terminologies on the percentage of participants who expect defection and cooperation respectively is shown in figure 7.5.b and summarised in table 7.7 for the first round. Notice that contrary to Bicchieri and Xiao [2009], our results do not indicate that satisfying condition 2(a) is a better predictor for actions than satisfying 2(b). However, our dataset is limited, as most participants satisfied 2(a) and 2(b) simultaneously. Figure 7.6.b shows how the terminologies affect participants' cooperation expectations throughout the ten rounds. Interestingly, we see that the gap between cooperation expectations in the three versions of the experiment diminishes with the number of rounds participants played. Indeed, it looks like participants' expectations in all three versions of the experiment might converge to a bit less than 50% expecting cooperation. Such convergence would be in accordance with participants' empirical observations of the random strategy they play against.

Summarising the results from the second experiment, we see that participants' choices depend on their social expectations. Further, we see that the terminologies change participants' social expectations across the ten rounds. Combining these results with the results from the first experiment, we see that the terminologies change participants' behaviour by changing their social expectations. This provides empirical support for the claim that two social norms are present in the prisoner's dilemma situation and that it is possible to change participants' behaviour in the situation by exposing them to empirical and normative cues equivalent to those found in microeconomics textbooks. Thus, the empirical results support the argument presented in chapter 6 that microeconomics textbooks can make their readers inclined to follow a social norm of (narrow) self-interest in economic situations.

Expectations first round	Collectivist terminology	Neutral terminology	Individualist terminology
Cooperate 2(a) only	0 (0%) [0]	3 (3.8%) [1]	7 (10.4%) [3]
Cooperate 2(b) only	2 (3.6%) [2]	4 (5.1%) [0]	4 (6.0%) [1]
Cooperate 2(a) and 2(b)	40 (72.7%) [38]	43 (55.1%) [36]	27 (40.3%) [23]
Defect 2(a) only	2 (3.6%) [0]	4 (5.1%) [4]	4 (6.0%) [3]
Defect 2(b) only	0 (0%) [0]	3 (3.8%) [2]	7 (10.4%) [4]
Defect 2(a) and 2(b)	13 (23.6%) [10]	28 (35.9%) [24]	29 (43.3%) [28]

Table 7.7: **Summary of participants' expectation in the first round.** The table summarises the expectations of the participants in the first round of the experiment across each of the three terminologies. The total number of participants in the collectivist version was 55, with 43 cooperating in the first round. The total number of participants in the neutral version was 78, with 41 cooperating in the first round. Finally, the total number of participants in the individualist version was 67 with 28 cooperating in the first round. Numbers in the soft parentheses indicate the percentage of participants for each terminology who had the given expectations. Numbers in the square brackets indicate how many of these participants acted in accordance with the expectations (e.g. cooperated or defected respectively). The symmetry in the table is caused by the fact that participants faced a binary choice.

7.4 Can social norms be stabilised in networks?

The results from the first two experiments support the claim that microeconomics textbooks can make their readers inclined to follow a social norm of (narrow) self-interest in economics situations by making them satisfy conditions 1 to 2(b). However, the results also show that people who are inclined to follow a social norm of narrow self-interest will only do so as long as they satisfy conditions 2(a) and 2(b). Thus, if readers of microeconomics textbooks experience sufficiently many people who do not follow a rule of (narrow) self-interest in economic situations then they may change their expectations such that 2(a) and 2(b) are no longer satisfied. In order to show that microeconomics can become self-fulfilling by promoting a social

norm of (narrow) self-interest, we thus also have to make probable that microeconomics teaching practices can *stabilise* a social norm of (narrow) self-interest, such that students will keep following the norm.

The aim of this section is to explore what will happen when people who are inclined to follow different social norms start interacting with each other in a network. This is done via a simulation experiment based on the data collected in the first experiment (section 7.2). In subsection 7.4.1, I present the theoretical set-up for the simulation experiment. In subsection 7.4.2, I report the results from the simulations and shortly discuss how they relate to the argument presented in section 6.4.[25]

7.4.1 Theoretical set-up

A simulation experiment in evolutionary game theory consists of running a series of simulations on a network. In order to do this, we have to choose a network, how the nodes interact in the network, and how the nodes update their actions.

Starting with the choice of network, we chose to conduct the simulation experiment on four non-directed networks in order to see how the network structures affect the final results:[26]

1 A two dimensional square lattice network, similar to the one used by Nowak and May [1992]. In this network, each node is fixed in a chequerboard with periodic boundary conditions and a von Neumann neighbourhood such that it can see four of its neighbours: up, down, left, and right.

2 A Barabási-Albert (BA) scale-free network, constructed by starting with a set number m of interconnected nodes, and then growing the network using a preferential attachment scheme where each new node

[25] For further discussion, see section 7.5.

[26] We also ran a preliminary study on an Erdos–Rényi network and a dynamic network constructed from interactions between students (measured by Bluetooth links) from the Copenhagen Networks Study [Stopczynski et al., 2014]. For further details, see Buchter et al. [2020, pp.25-34].

is attached to m existing nodes with probabilities proportional to the degrees of the nodes [Barabási and Albert, 1999]. In evolutionary game theory, BA networks are typically constructed with parameters ranging from $m = 4$ to $m = 8$ [Santos and Pacheco, 2005, Wu et al., 2007]. Because of this, we chose $m = 6$.

3 A text message (SMS) network, constructed using data from one month of the Copenhagen Network Study which collected information on the interactions between 1000 students at the Technical University of Denmark in the course of a year [Stopczynski et al., 2014]. The SMS network contains $n = 457$ nodes, where the nodes are connected if one of the people they represent has texted another during the one-month observation period.

4 A Facebook network, constructed from the same one-month period of the Copenhagen Network Study. The Facebook network consists of $n = 800$ nodes, where nodes are linked if the people they represent are friends on Facebook.

The network structures are illustrated in figure 7.8. The lattice and BA networks are used for comparability to previous results in the simulation literature on evolutionary game theory [Nowak and May, 1992, Barabási and Albert, 1999, Santos and Pacheco, 2005, Du et al., 2008]. We constructed the SMS and Facebook networks in order to test the collective effects of students who follow different social norms interacting with each other. Since the networks are based on real student interactions, they approximate the social networks of economics students discussed in section 6.4.

In alignment with the experiment, we chose that the nodes should interact by playing a *prisoner's dilemma game* (PD). However, in order for the simulation results to be comparable to the current simulation literature on PD, the payout matrix of the PD played in the simulations was changed from a *strong* PD with payouts $T = t$, $R = 2$, $P = 1$, and $S = 0$ where $t \in (2, 4)$ to a *weak* PD with payouts $T = t$, $R = 1$, and $P = S = 0$ where $t \in [1, 2]$ [Nowak and May, 1992, Platkowski, 2009, Grujić et al., 2010, Gracia-Lázaro et al., 2012, Holme et al.,

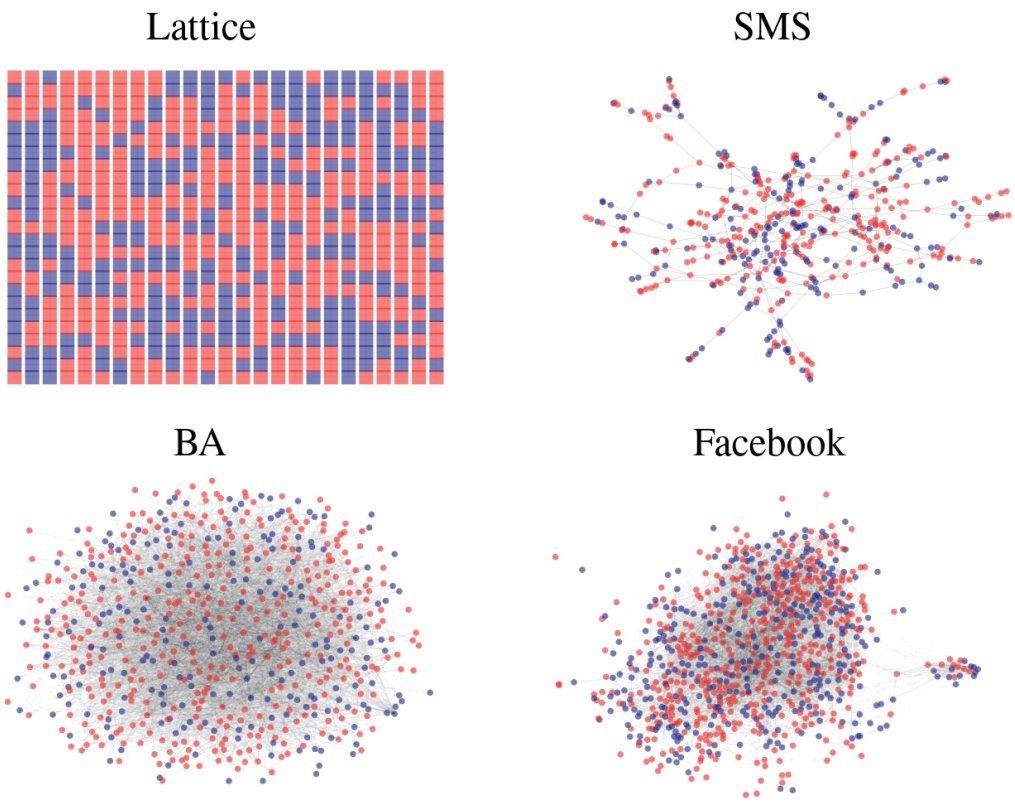


Figure 7.8: **Illustration of the four network structures we consider.** Examples of the four network structures taken from our preliminary network analyses. The red dots indicate nodes that defect while blue dots indicate nodes that cooperate.

2003].[27] The reason this is done in the literature, is that simulation heuristics typically only have payouts as their input. Because of this, a *strong* PD will favour defection such that once a node starts defecting, it will never change its strategy. This typically leads to a simulation outcome where all nodes defect. Using a *weak* PD, however, makes it equally attractive for a node (whose opponents play defect) to play defect and cooperate. This means that the system can oscillate between cooperation and defection, which makes it possible to study different effects caused by the network structure or small changes in the nodes' decision heuristics.

[27] This is not a trivial change since it changes the strategic incentives of the game. Though it is worth worrying about, I am going to bracket the concern in order to work within the practices established in the literature.

Finally, we chose a model for how nodes should update their strategies. In order for the model to correspond to the design of the first experiment, we chose to make each node update their strategies via a softmax function that determines their probabilities for cooperating (p_c) and defecting (p_d) based on their own payoff and the payoffs of one random neighbour [Wu et al., 2007, Blume, 1993]:

$$\begin{aligned} p_c &= \frac{e^{\beta \mathbf{w}_c \cdot \mathbf{x}}}{e^{\beta \mathbf{w}_c \cdot \mathbf{x}} + e^{\beta \mathbf{w}_d \cdot \mathbf{x}}}, \\ p_d &= \frac{e^{\beta \mathbf{w}_d \cdot \mathbf{x}}}{e^{\beta \mathbf{w}_c \cdot \mathbf{x}} + e^{\beta \mathbf{w}_d \cdot \mathbf{x}}} \end{aligned} \tag{7.1}$$

$$\begin{aligned} p_d &= \frac{1}{1 + e^{-\beta \mathbf{w} \cdot \mathbf{x}}}, \\ p_c &= 1 - p_d, \\ \mathbf{w} &= \mathbf{w}_d - \mathbf{w}_c. \end{aligned} \tag{7.2}$$

Here, $\mathbf{x}$ is an input vector which contains the information the node uses to make their decision. $\mathbf{w_i}$ is a weight vector that assigns weights to each entry in $\mathbf{x}$, where $i = c$ indicates that the weights are associated with a strategy of cooperation and $i = d$ indicates that the weights are associated with a strategy of defection. Thus, $\mathbf{w_i}$ represents the real-valued relative importance of each of the entries in $\mathbf{x}$. Since we wanted to allow for biases in the node's decision, we introduced a bias term, b_i, in $\mathbf{w_i}$ and designed $\mathbf{x}$ such that it would always be a part of $\mathbf{w_i} \cdot \mathbf{x}$. Finally, β indicates how stochastic the node's choice is. The larger the value of β, the more the node's choice will depend on the values given by $\mathbf{w_i} \cdot \mathbf{x}$. Thus, as β goes to infinity, the node will deterministically choose the strategy with the highest $\mathbf{w_i} \cdot \mathbf{x}$. If $\beta = 0$, the node will choose its strategy uniformly at random independent of the values of $\mathbf{w_i} \cdot \mathbf{x}$.

In order to increase the realism of the update heuristic, we fitted the model to data from the first experiment (excluding economists) for each of the three terminologies. The best fit was achieved by using an input vector, $\mathbf{x}$, with the

	w_{sd}	w_{sc}	w_{nd}	w_{nc}	b_{cc}	b_{cd}	b_{dc}	b_{dd}	β
C	.31	.28	.017	-.042	-.81	-.42	.00	-.02	3.5
N	.47	.36	-.01	-.03	-.78	-.42	.052	.047	3.0
I	.34	.31	.015	-.035	-.75	-.25	.17	.037	2.7

Table 7.9: **Best parameter fits under the three different terminologies; C, N, and I.**

following entries:[28]

$$\mathbf{x} = (\delta_{H,cc}, \delta_{H,cd}, \delta_{H,dc}, \delta_{H,dd}, \delta_{s,d} \cdot p_s, \delta_{s,c} \cdot p_s, \delta_{n,d} \cdot p_n, \delta_{n,c} \cdot p_n)' \tag{7.3}$$

Here, H (history) indicates what strategy the node and the node's opponent played in the last round of the game, with $\delta_{H,ij}$ being a Kronecker delta which is one if the node and their opponent played strategies i and j in the previous round and zero otherwise. Further, the payout of the node is included by s (self), with p being the payout, and the payout of the random neighbour is included as n (neighbour) with p being the payout. In this model, we allowed the bias term, b_i, to vary depending on the strategies of the node and the node's opponent in the last round. Thus, the weight vector consisted of eight entries: four bias terms (b_{cc}, b_{cd}, b_{dc}, and b_{dd}) corresponding to the immediate history of the node's last game, and four weights (w_{sd}, w_{sc}, w_{nd}, and w_{nc}) corresponding to the weights assigned to payoffs received by the node and the node's neighbour.[29] The parameters in the weight vector that best fitted data for each of the three versions of the experiment are presented in table 7.9.

Next, we made three adjustments to the fitted model in order to make it correspond to the models found in the simulation literature on evolutionary game theory. First, we shifted the bias terms, so that biases for the model with neutral terminology was zero, while the differences between the bias

[28] In order to get the best fit to data, we tried to fit the model with several different version of $\mathbf{x}$. A description of the different models we tried to fit along with a summary of how well they fitted (using their log-likelihoods, the accuracy of the models, their AIC scores, and their $F1$ scores) is reported in Buchter et al. [2020, pp.13-14].

[29] Notice that the value of t is implicitly a part of the payoffs for the node and its neighbour.

terms for the three versions of the experiment were retained.[30] Second, the stochasticity in the fitted model implied that clusters of agents who act similarly are less stable. This negates interesting network effects [Gracia-Lázaro et al., 2012], wherefore we chose to change the update heuristic of the fitted model to an *individual softmax heuristic*, by including the rule that if the node and its random neighbour played the same strategy, then the node will keep playing that strategy.[31] This change fits well with Bicchieri's account of social norms, where an agent would not have incentive to diverge from a social norm, if the agent observes that other people in their reference network also follow the norm. Third, we found that participants in our experiment were not very influenced by the choice and payoff of the random neighbour they were informed about. This is in accordance with what was observed by Grujić et al. [2010]. Since it affects the simulation results adversely if the nodes do not pay attention to their neighbours [Gracia-Lázaro et al., 2012], we changed the weights in the weight vector such that the node will use the same weight for its neighbour's payout as for its own payout. This matches the literature, in which every individual heuristic we encountered also treats payout of the node and the payout of its neighbour as equally important. The change also makes sense in the social norm framework we consider, since it supports the idea that the node's decision is conditional on other nodes' behaviour. Given the three adjustments, table 7.10 presents the final parameters used in the simulation experiment for the three terminologies.

In the simulation experiment we conducted, we wanted to consider how nodes that are inclined to follow different social norms interact in a network. To do this, we defined a parameter, ρ_I, denoting the fraction of nodes in a given simulation that uses a decision heuristic based on the individualist model. The remaining $1 - \rho_I$ nodes in the network are assigned the collectivist model. Thus, we did not use the neutral model in our simulation experiment.

[30] We did this because the fitted model displayed a strong bias toward defection - possibly because participants in the experiment had played a *strong* PD. For details on how we made the adjustments see Buchter et al. [2020, pp.14-15].

[31] We did not make this change before we fitted the model to the data, since some of the participants in the experiment have acted contrary to the deterministic rule, making the likelihood of having this observation under the model zero.

	w_{sd}	w_{sc}	w_{nd}	w_{nc}	b_{cc}	b_{cd}	b_{dc}	b_{dd}	β
C	0.23	0.21	0.23	0.21	-0.11	-0.049	-0.033	-0.046	4.6
N	0.26	0.24	0.26	0.24	0	0	0	0	3.9
I	0.22	0.17	0.22	0.17	0.054	0.1	0.054	-0.0072	5.7

Table 7.10: **Adjusted parameters for the agent heuristic.**

In all simulations we ran, 10% of the nodes updated their strategy per round. The reason for this is to avoid artificial effects that can occur if all nodes update their strategies in every round [Tomassini et al., 2007, Newth and Cornforth, 2009, Grilo and Correia, 2009].

Finally, we defined two measures to determine how a strategy fared in the simulations. The first measure, *pervasiveness*, determines how widespread a certain strategy is in the entire network. It is defined as the fraction of nodes in the network that chooses a given strategy (cooperate or defect) by the end of the simulation. Thus, pervasiveness can be seen as a global measure of how much a given strategy is adopted in the network. The second measure, *prevalence*, determines how tightly nodes with a given strategy are clustered together in the network. It is defined as the z-score of the observed number of links between nodes with the same strategy compare to a random permutation null model. Thus, the higher the prevalence number, the more clustered the nodes with a certain strategy are, compared to a random distribution of these nodes. The prevalence measure can be seen as a local measure of how much nodes with a certain strategy "stick together".

7.4.2 Simulation results

We ran several simulations for each of the four network structures, with 10^4 rounds in each simulation. In the simulations we used different parameters of ρ_I and t, to see how the two effects interact. The results from the simulations conducted on the Facebook network are summarised in figure 7.11 [Buchter et al., 2020, Fig.3, p.8].

Figure 7.11.a shows how likely it is for cooperation to die out in the simulations for combinations of $\rho_I \in [0, 1]$ and $t \in [1, 2]$. For each cell in the

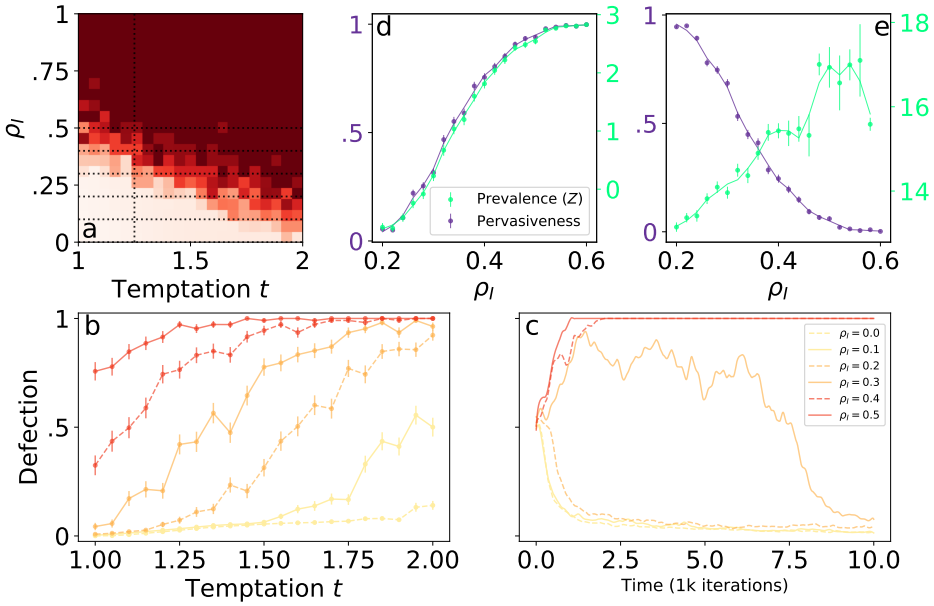


Figure 7.11: **Simulation results for the FB network for a range of values of t and ρ_I.** **a** The tendency of cooperation to disappear entirely in the simulation, for combinations $\rho_I \in [0, 1]$ and $t \in [1, 2]$. The darker the cell, the higher the fraction of ten simulations ending in a state where no node cooperated after 10^4 rounds. **b** The mean defection rates for selected values of ρ_I (indicated with dashed lines in **a**) - averaged over the final 5000 iterations of 100 simulations - as a function of t. **c** Progression of one simulation for specific values of ρ_I and $t = 1.25$ (indicated by the intersections of lines in **a**). The lines are smoothened using a Savitzky–Golay filter with window length 400 and polynomial order five. **d** The development of *pervasiveness* and *prevalence* for defection for different values of ρ_I. The analysis is conducted for $t = 1.25$ (as indicated by the vertical line in **a**). **e** The development of *pervasiveness* and *prevalence* for cooperation for different values of ρ_I in for $t = 1.25$.

figure, we ran ten simulations. The colour indicates the fraction of these simulations where cooperation died out, with a darker shade representing an increased fraction of simulations where this happened. The figure shows that increasing t for a given ρ_I will gradually increase the number of simulations where cooperation dies out. However, variations in ρ_I for a given t can completely change whether cooperation survives as a strategy in the network. Thus, the effect of increasing the number of nodes following the individualist model is much larger than the effect of increasing the monetary incentive to defect.

Looking at specific values of ρ_I (indicated by the horizontal dotted lines in figure 7.11.a), figure 7.11.b confirms this, and shows how changes in the value of ρ_I for different values of t can take the entire simulation across a tipping point, between complete (or almost complete) cooperation to complete defection.

Figure 7.11.c displays the fraction of defection in the network for $t = 1.25$ (indicated by the vertical dotted line in figure 7.11.a) during the 10^4 rounds of the simulations. This is displayed for the selected values of ρ_I also used in figure 7.11.b. The figure shows that for high values of ρ_I, the entire network quickly starts defecting, while the entire network quickly starts cooperating for low values of ρ_I. For values around $\rho_I = 0.3$, the system is volatile and can go from complete defection to complete cooperation.

Next, we considered our two measures *pervasiveness* and *prevalence*. Figure 7.11.d shows the pervasiveness and prevalence of defection for different values of ρ_I for $t = 1.25$. As ρ_I increases, the fraction of nodes that defect - and thus the pervasiveness of defection - is increased. Further, as ρ_I increases, defectors become more clustered together, increasing the prevalence of defection. This also means that for networks with a low ρ_I, where most nodes will tend to cooperate, the nodes who defect will typically not be linked to other nodes that defect (which makes sense given the payoff structure of PD). Figure 7.11.e summaries the same results for cooperation. Here we see that the fraction of nodes that cooperate in the network decreases as ρ_I increases, thus decreasing the pervasiveness of cooperation. Further, we see that increases in ρ_I lead to an increase in prevalence for cooperation. This

means that as defection becomes the predominant strategy in the network, the nodes that keeps cooperating will typically be clustered together with other nodes that cooperate. Thus, cooperation is more likely to survive as a strategy, if the nodes that cooperate are linked to each other.[32]

Finally, we examined the effects of using different network structures. Figure 7.12 shows how the effects of ρ_I and t differ between the SMS, Facebook, BA, and square lattice (SL) networks. The columns in figure 7.12 display the same figures as the ones shown in figure 7.11.a-c for the Facebook network. The figure shows that the network structure does have an impact on how likely it is that cooperation will survive as a strategy in the simulations. Specifically, we see that this is the least likely in the SMS network and most likely in the square lattice network. The simulation results in the Facebook network and BA network are relatively similar.

Summarising the findings from the simulation experiment, we see that the distribution of nodes modelled on behaviour of people exposed to the individualist terminology and nodes modelled on the behaviour of people exposed to the collectivist terminology can completely tip which strategy will end up being the predominant strategy in a network. These results are in accordance with the argument by Nyborg et al. [2016] that individual behavioural differences caused by social norms may drive a collective system across a tipping point and cause dramatic collective effects. Further, though network structures do affect the simulation results, our findings are robust across the different artificial and real networks considered. This supports the argument presented in section 6.4 that interaction among microeconomics university students can stabilise a social norm of (narrow) self-interest in their social network.[33] Finally, we see that the collective effect of exposure to different terminologies are larger than the effect of t. This finding indicates the importance of social norms in determining our behaviour. It also supports Bicchieri's [2016, pp.153-156] assessment that monetary incentives may not be an effective tool to change behaviour in situations involving social norms.

[32] This makes sense since cooperating nodes that are surrounded by defecting nodes will receive the lowest possible output and therefore change strategy.

[33] Notice that the nodes in the simulation experiment use a soft threshold model as their update heuristic rather than a normal threshold model, as discussed in sections 6.4 and 6.2.3.

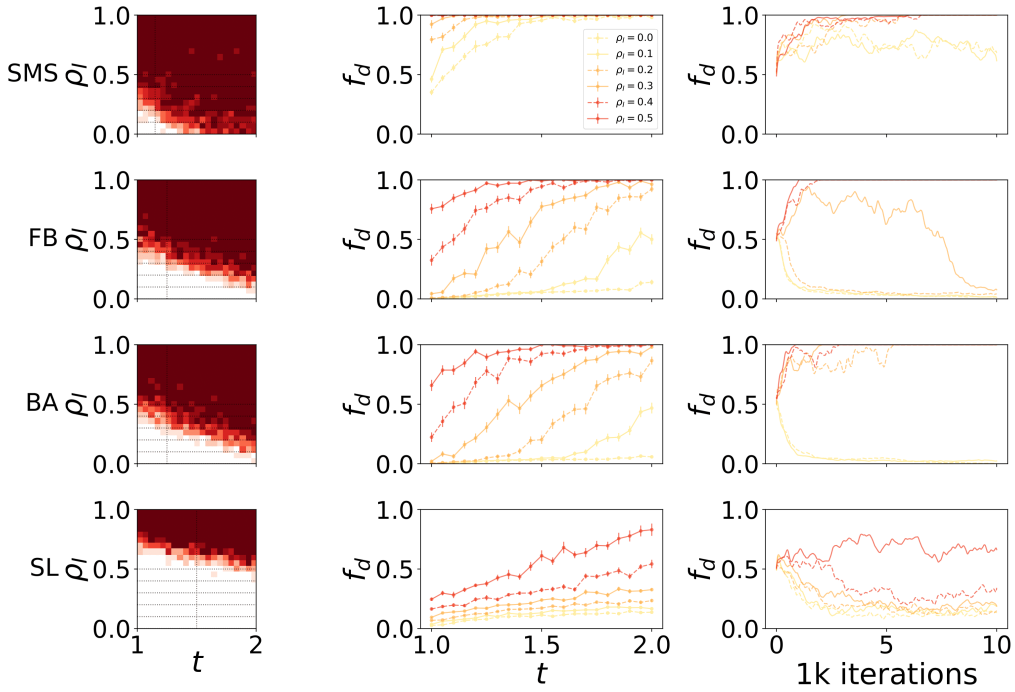


Figure 7.12: **The effect of network structure on simulation results.** Each row corresponds to a given network structure: SMS, Facebook, BA, and square lattice. **The left column** shows the tendency of cooperation to disappear in simulations for combinations of $\rho_I \in [0, 1]$ and $t \in [1, 2]$. The darker the cell, the higher a fraction of ten simulations resulted in a state where no nodes cooperated after 10^4 rounds. **The middle column** shows the mean fraction of defection averaged over the last 5000 rounds in the simulations as a function of t, for selected values of ρ_I. Each point represents ten simulations for the SMS and square lattice networks, and 100 simulations for the Facebook and BA networks. The selected values of ρ_I and t are indicated in the left column figures by dotted lines. **The right column** shows the fraction of defection for one simulation as a function of rounds played in the simulation for a given value of t and selected values of ρ_I (as indicated by the dotted lines in the left column figures). The lines are smoothened using a Savitzky–Golay filter with window length 400 and polynomial order five.

Finally, the finding is contrary to the assumption found in microeconomics textbooks that individuals are only motivated by their narrow self-interest in situations involving a possibility for monetary gains.

7.5 Discussion

The aim of the three experiments described in this chapter is to test whether the arguments I provided in chapter 6 are supported by empirical observations.

The first two experiments were designed to test whether exposure to microeconomic terminology can make people follow a social norm of defection or cooperation. The first experiment confirmed that participants' choices were conditional on which terminology they were exposed to and that participants exposed to the individualist terminology defected more than participants in the control group while participants exposed to the collectivist terminology cooperated more than participants in the control group. The experiment also showed that economists in general defected more than non-economists and that this difference is comparable to the increase in defection for non-economists between the neutral terminology and the individualist terminology. The second experiment showed that participants' choices are correlated with their expectations, such that participants defect if they expect 2(a) and 2(b) to be satisfied for defection and cooperate if they expect 2(a) and 2(b) to be satisfied for cooperation. The experiment also showed that the individualist terminology increased the number of participants who satisfied 2(a) and 2(b) for defection while the collectivist terminology increased the number of participants who satisfied 2(a) and 2(b) for cooperation. Taken together, the two experiments correspond to the two measures, Bicchieri [2016, ch.2] proposes to measure the existence and causal power of a social norm. The first experiment supports that condition 2 is satisfied for a social norm of defection (narrow self-interest) and for a social norm of cooperation. The second experiment shows that the terminologies can change participants' social expectations, making them more inclined to follow one social norm rather than the other. Thus, we can conclude that participants' behaviour in a prisoner's dilemma game is affected by two

social norms and that exposing participants to different terminologies can make them more inclined to follow one social norm rather than the other. This supports the argument presented in section 6.3.

The findings also help us explain the behaviour of economists observed in the first experiment.[34] Since economists have already been exposed to microeconomics textbooks, a social norm of narrow self-interest may already be salient to them. According to Bicchieri [2006, p.48], we are likely to apply the norms we already know and are familiar with when we are faced with new situations. Thus, one explanation why economists defect more is that they are already inclined to follow a social norm of narrow self-interest. This explanation is also supported by the results of Frey and Meier [2003], Cappelen et al. [2015], and Gerlach [2017], who all find that one reason economists behave differently is that they expect their peers to behave in that way.

Finally, the findings help us understand why the effects of the terminologies remain during the ten rounds of the game even though participants play against a strategy chosen at random. When a social norm is activated, Bicchieri [2006, p.97] suggests that people will tend to follow it by default. Further, they will need sufficient evidence in order to change their expectations such that 2(a) and 2(b) are no longer satisfied. This can explain why participants' choices still vary between the three versions of the experiment by the end of the tenth round.

The simulation experiment was designed to test which collective effects that can emerge when people who are inclined to follow different social norms start interacting in a social network. Thus, the simulation experiment tests how participants might have chosen, if they interacted with each other (rather than with a random strategy) in different networks. We tested this effect in a separate experiment in order to distinguish between the effect of terminologies and the effect of other people's choices. In order to make the simulation experiment as realistic as possible, we based the nodes' update heuristics on observed behaviour and used social networks constructed from

[34]The finding that economists defect more than non-economists is consistent with the results from previous studies reported in chapter 3.

the interactions among university students. The results show that the fraction of people inclined to follow a social norm of defection can tip the outcome of the entire collective system regardless of the monetary incentives to defect. Though the situation in the simulation experiment does not correspond directly to the argument presented in sections 6.4, it none the less does provide suggestive evidence that if sufficiently many people who are inclined to follow a social norm of narrow self-interest interact in a social network then this can cause the entire network to follow this norm. Thus, it supports the claim that microeconomics teaching practices can help stabilise a social norm of (narrow) self-interest by letting the students inclined to follow this norm interact with each other.

I end this section by discussing three potential worries about the experiments. The discussions will also include some considerations on the limits of our experiments as well as suggestions to possible future research.

First worry: the terminologies are too obvious

As described in section 7.2.1, the individualist and collectivist terminologies consist of two parts. The first is a framing part: using "your opponent" and "your co-player" respectively. This was used to provide the participants with *empirical cues* about the situation. The second part is the introduction of a microeconomic concept which participants are asked to apply to the strategies in PD. The concept was designed to provide participants with *normative cues* about the situation. Thus, the two parts of the terminologies were intended to affect participants empirical and normative expectations, respectively.

The literature from chapter 3, however, suggests that framing a situation as competitive or cooperative can have an effect on behaviour [Kay and Ross, 2003, Kay et al., 2004, Engel and Rand, 2014].[35] It also shows that the introduction of a microeconomic text excerpt can cause changes in participants' behaviour [Ifcher and Zarghamee, 2018]. This raises the concern that by providing two cues that can have a behavioural effect, we are pushing

[35]Note that Belaus et al. [2018] were not able to replicate Kay and Ross [2003].

the participants rather than nudging them.

There are two reasons why we none the less included both types of stimuli in the two experiments. First, the aim of the experiments is to test whether conditions 2, 2(a), and 2(b) are satisfied for the participants. In order to do this, we needed both empirical and normative stimuli. Second, as I have argued in sections 5.4 and 6.3, microeconomics textbooks contain both types of stimuli with regard to economic situations. Thus, the stimuli we use in the experiment are in accordance with and not stronger than the ones present in microeconomics textbooks.

A related worry is that we obtained the results that we did because the terminologies "told" the participants what they were expected to do. This concern is related to the effect that participants in experiments often try to guess the purpose of the experiment in order to act accordingly [Nichols and Maner, 2008]. Since the terminologies use several stimuli, they may have provided participants with sufficient information to guess what they were expected to do.

In order to accommodate this worry, we ended the experiment by asking participants whether they had guessed the purpose of the study. Table 7.13 summarises the fraction of participants with relevant guesses for each data source in the two experiments. Participants in the column "Effect of language" guessed that we tested how language influenced people's choices. Thus, these participants guessed that we tried to manipulate choices through the terminologies. Participants in the column "We should defect" stated that they thought the right thing to do was to defect in the situation (many mentioned that the game was a prisoner's dilemma game). Participants in the column "We should cooperate" stated that they thought the right thing to do in the game was to cooperate. The responses in the last two columns generally consist of participants' own attitudes towards the game or repetition of the aim we told them in the beginning of the experiment (to study whether people are rational or to study whether people cooperate). While the last two columns need not express that participants have guessed the aim of the experiment, we have chosen to err on the side of caution and conduct a follow-up analysis that excludes all participants whose guesses

Experiment	Terminology	"Effect of language"	"We should defect"	"We should cooperate"
First experiment BRL	Individualist	4%	14%	2%
	Neutral	0%	8%	4%
	Collectivist	4%	9%	5%
First experiment MTurk	Individualist	2%	3%	2%
	Neutral	1%	2%	1%
	Collectivist	1%	2%	0%
Second experiment	Individualist	1%	0%	0%
	Neutral	0%	0%	0%
	Collectivist	4%	0%	2%

Table 7.13: **Summary of the fraction of participants with relevant guesses on the purpose of the experiments.**

are summarised in the table.

Making a follow-up analysis on the main findings, we see that all results remain significant. Considering the data set in the first experiment consisting of MTurk participants and BRL participants excluding economists, we still see that the terminologies have a significant impact on participants' choices: $p = 1.4 \cdot 10^{-8}$, $H = 36$, Kruskal-Wallis test. Participants defect significantly more in the individualist version compared to the neutral version ($p = 9.34 \cdot 10^{-6}$, $t = 4.31$, one-tailed Conover-Iman post hoc analysis) and cooperate significantly more in the collectivist version compared to the neutral version ($p = 0.03$, $t = 1.87$, one-tailed). The same results are significant for economists in the BRL data set and for the two full data sets combined. In the second experiment, we still see that participants' choices are correlated with their expectations ($p = 1.6 \cdot 10^{-24}$, $z = 10.2$, proportions z-test). The follow-up analysis also shows that the terminologies affect how many people satisfy 2(a) and 2(b) with respect to defection and cooperation respectively ($p = 0.043$, $H = 6.3$ and $p = 0.0052$, $H = 11$, Kruskal-Wallis test). Thus, the results of the experiments are confirmed even when all participants who may have tried to act in accordance with what they were "told" are excluded.

Second worry: are there alternative explanations for the observed behaviour?

Combining the first two experiments, we concluded that two social norms exist in a prisoner's dilemma situation and that the terminologies influence behaviour by influencing whether participants satisfy conditions 2(a) and 2(b) for defection or cooperation. One may, however, ask whether there are alternative explanations for the behavioural changes that are compatible with data. Here, I consider five alternative explanations.[36]

The first alternative explanation is that the different terminologies may have enabled participants - who already wished to maximise collective or individual wealth - to do this by making sure that they understood how to do it. According to this explanation, the terminologies affect behaviour by teaching participants how the game works, so that participants are able to act in accordance with their own ends. The second alternative explanation is that the terminologies may have changed participants' preferences [Cappelen et al., 2015], so that the number of altruistic participants increases in the collectivist version of the experiment and the number of narrowly self-interested participants increases in the individualist version of the experiment. However, if either of these alternative explanations are correct, then we would *not* expect the participants' behaviour to depend on their expectations about what other participants do. This is because each person will always maximise individual wealth by choosing to defect (regardless of what the other player does). Similarly, each person will always maximise collective wealth by choosing to cooperate (regardless of what the other player does).[37] Since the second experiment shows that there is a strong correlation between participants' expectations and what they choose to do (across all three versions of the experiment), the two alternative explanations are not supported by data.

A third alternative explanation is that the difference in behaviour can be

[36] These are the five alternative explanations that we have been asked about when presenting the results of our experiments at various seminars.

[37] The control questions in each version of the experiments ensured that participants knew this.

explained by a false consensus effect. A false consensus effect is when people project their own beliefs and values on to others [Ross et al., 1977]. In the case of our experiments, this would mean that participants in the collectivist version, for example, would cooperate because they believed it was the best thing to do, and therefore believed that other participants would believe the same. Note that this explanation does not explain *why* participants come to change their values when exposed to different terminologies. Further, if the initial difference between the versions were due to a false consensus effect, then we would expect it to disappear once participants started observing the random strategy they played against. Since the behavioural difference between the three versions is maintained throughout the ten rounds, a false consensus effect cannot explain our results.

A fourth explanation is that differences in participants' behaviour are caused by an anchoring effect. An anchoring effect occurs if people receive information which they then use as a reference point when making a judgement [Furnham and Boo, 2011]. Such informational anchors can be used as a coordination device when trying to coordinate with other people. In the case of our experiments, one might think that the reason participants, for example, cooperate more in the collectivist version is that they take the collectivist terminology as an anchor they can use for coordination. This psychological explanation is not something we tested for in our experiment. However, it is consistent with Bicchieri's theory of social norms as the theory does not tell us what motivates people to conform to a social norm. It may be fear of sanction, a desire to be praised, a wish to coordinate, or some other emotion that motivates norm compliance [Bicchieri, 2006, p.42]. Whatever motivation it is, it is consistent with the presence of social norms, if the definition for social norms is satisfied, as it is in our case.

Finally, participants might change their behaviour simply because they want to do what is expected of them, and the terminologies indicate what this is. This explanation has two versions, depending on how we interpret the scope of *whose* expectations participants want to conform to. If participants conformed to the expectations of the experimenters or of game theory, then their behaviour would not be strongly correlated with their social expectations

regarding other players. Thus, the results of the second experiment indicate that this is not the case. If, however, participants want to conform to the social expectations of other participants, then this explanation is in alignment with what we observe and with the theory that participants' behaviour is motivated by the presence of two social norms.

Third worry: what can we conclude from the simulation experiment?

Finally, we made several adjustments in the simulation experiment in order to conform with the current practice in the simulation literature on evolutionary game theory. Because of these adjustments, the simulation results cannot be directly assumed to hold in the real world. Instead, they show that under some specific circumstances, the adjusted behavioural changes caused by the terminologies in the first experiment can lead an entire network across a tipping point.

Further, the simulation experiment does not correspond directly to the argument presented in section 6.4 for how a social norm of (narrow) self-interest can be stabilised via microeconomics teaching practices. Still, it shows that if sufficiently many people who are inclined to follow a social norm of narrow self-interest repeatedly interact with each other in a social network - constructed from student interactions - then this can stabilise the norm in the network. Thus, it lends some support to the plausibility of the argument.

Since the first two experiments did not allow participants to play against each other, an obvious suggestion for future research is to let participants exposed to the same terminology play against each other in one of the network structures studied in the simulation experiment. By conducting this study, we would be able to see how playing against other participants will change the cooperation/defection rate in the three versions throughout the ten rounds of the experiment. In a second study, one could then let participants exposed to different terminologies play against each other in the same network structures. This would provide us with data on real behaviour that can be compared to the simulation results presented in section 7.4. Conducting these two follow-up experiments would contribute both

to the literature on social norms and to the literature on evolutionary game theory.

7.6 Conclusion

In this chapter, I have reported the results of two laboratory experiments and one simulation experiment conducted in order to empirically test the arguments presented in chapter 6.

The findings of the first two experiments show that participants playing prisoner's dilemma either follow a social norm of defection (narrow self-interest) or of cooperation. They also show that terminologies containing the same stimuli as microeconomics textbooks can change participants' behaviour by changing their normative and empirical expectations with regard to these two norms. Thus, the experiments support the argument that microeconomics textbooks can make their readers inclined to follow a social norm of (narrow) self-interest in economic situations.

The results from the simulation experiment further show that the fraction of nodes in a social network that act in accordance with participants who are inclined to follow a social norm of narrow self-interest can tip the entire network from cooperation to defection. This lends some support to the argument that a social norm of (narrow) self-interest can be stabilised in a social network via microeconomics teaching practices.

Combining the arguments in chapter 6 with the empirical results in this chapter, I conclude that one way in which microeconomics can become self-fulfilling is by promoting a social norm of (narrow) self-interest in economic situations via the dissemination of microeconomics textbooks and teaching practices. As a part of the outlook in the next chapter, I discuss how we can change microeconomics textbooks and teaching practices if we want to avoid promoting this norm.

CHAPTER 8

CONCLUSION

In this thesis, I have argued that microeconomics textbooks and teaching practices can promote a social norm of self-interest - and often narrow self-interest - in economic situations. Thus, microeconomics is a self-fulfilling science that can have social implications for our societies by changing people's behaviour. My analyses also show that, even though microeconomics textbooks seek to maintain a distinction between positive and normative economics, they are not able to do so; as simply using value laden words like *rational* can affect people's behaviour. Thus, the thesis also provides evidence that the distinction between positive and normative economics may not be as easily maintained as Friedman [1953] suggests.

The arguments in this thesis contribute to at least three areas of philosophy. First, to philosophy of science and the discussion of self-fulfilling sciences by showing one mechanism through which scientific descriptions can become self-fulfilling. This fills a gap in the literature, since no one previously - to my knowledge - has presented a satisfactory discussion of *how* the phenomenon can occur for microeconomic descriptions. Second, this thesis contributes to the discussions in social ontology on social norms by providing evidence that behaviour in economic situations is guided by the presence of different social norms. Third, the thesis contributes to philosophy of economics by showing how microeconomics textbooks are unable to maintain the divide

between positive and normative economics as recommended by Friedman [1953]. In line with the debate in philosophy of science on values in science, it thus shows that microeconomics may not be as value free as it is often claimed to be. This makes it important to discuss the values promoted by microeconomics textbooks, and teaching practices, and to determine whether the behavioural changes they can cause are desirable.

Finally, the arguments in this thesis can have real life consequences if we do not think it is desirable for a social norm of (narrow) self-interest to be promoted in economic situations. By showing that the behavioural changes caused by microeconomics textbooks and teaching practices are due to a social norm, the arguments make it possible to intervene with the current practices and prevent any further promotion of this norm. I end this thesis, first, by summarising its arguments and, second, by discussing some preliminary suggestions on how to mitigate the consequences of microeconomics textbooks and teaching practices.

Summary of the thesis

In philosophy of science, the phenomenon that scientific descriptions can have unintended effects has been discussed under the names of - for example - *self-fulfilling prophecies*, *the Oedipus effect*, *self-fulfilling predictions*, *reflexivity*, *looping effects*, and *performativity*. By analysing the definitions and uses of the different concepts, I have shown that there are subtle differences both between each concept and between the meaning different theorists attach to the same concept. Analysing the problems related to the phenomenon we see that these fall in three categories, and that different features of the phenomenon become relevant depending on which problem is discussed. To avoid confusion, I therefore suggest to explicitly state the features we are concerned with along seven dimensions. Further, I suggest using the different concepts already present in the literature to indicate which problem we consider. Specifically, I argue for using the term *self-fulfilling science* to describe a situation where the dissemination of scientific descriptions aimed at all agents can have social implications by influencing behaviour.

Looking at the literature concerned with self-fulfilling science, we see that very little has been written with respect to *how* this phenomenon can occur. Thus, the aim of this thesis is to fill a gap in the philosophy of science literature by presenting a thorough argument for one way scientific descriptions can become self-fulfilling. I do this by looking at the case of microeconomics and argue that dissemination of the rationality assumption as it is presented in microeconomics textbooks can - along with microeconomics teaching practices at universities - promote a social norm of self-interest and often narrow self-interest in economic situations.

In order to make this argument, I first reviewed the current empirical findings comparing the behaviour of economists and non-economists in economic situations. The results show that there is a behavioural difference between the two groups and that it can be caused by the exposure to microeconomic text excerpts describing the rationality assumption or by framing a situation as an economic or competitive situation. The results also show that the difference between economists and non-economists disappears in situations where there are salient non-economic concerns such as the concern for keeping a promise or the concern of helping people who are worse off due to no fault of their own. Based on these findings, I conclude that microeconomics *is* a self-fulfilling science and that this effect can be caused by exposure to the rationality assumption as it is presented in microeconomics textbooks.

The claim that microeconomics can cause behavioural changes may be disputed due to the methodology presented by Friedman [1953]. Here, Friedman argues that it is possible to distinguish between positive and normative economics such that positive economics is aimed at making predictions about how the world is while normative economics is concerned with how the world ought to be. According to Friedman, the former is independent of the latter such that there are no normative judgements in positive economics. If this is correct, and since microeconomics textbooks are primarily concerned with positive economics, it is unclear how microeconomics textbook will be able to influence behaviour. Thus, I next turned to the question of whether positive microeconomics is indeed a value free science.

Looking at the historical development of economics, we see that it is closely related to the philosophical theories in 18th century that human nature is controlled by passions. Using the cases of Mandeville, Smith, Hume, and Voltaire, I argued how self-interest and greed (or narrow self-interest) has been promoted in the early classical economic literature as both publicly and individually beneficial. Further, I argued that these normative arguments for the benefits of self-interest can also be found in recent neoclassical literature exemplified by the writings of Hayek and Friedman. Contrary to Friedman's claim, we thus see a close historical connection between positive microeconomics and the normative claims that greed and self-interest are beneficial. The analyses also show that the normative arguments fail to establish that greed and self-interest actually are beneficial. This raises the questions why the assumption that people act self-interestedly is still a part of contemporary microeconomics and how the assumption is presented in textbooks today.

By analysing the development of the rationality assumption, we see that changes to the description of human behaviour in economics is closely linked to changes in economic theory itself. First, Mill's focus on humans' desire to increase their wealth in economic situations helped him differentiate economics from other social sciences. Second, Jevons's description of people as always striving to maximise their hedonistic utility enabled him to use mathematical models to analyse behaviour in economic situations. Thus, the rationality assumption is persistent in microeconomic theory because it played a crucial role in making microeconomics the science it is today. Further, it is present in all microeconomic models because it has the dual function of being both an economic assumption about human behaviour in economic situations and a formal assumption that sets mathematical constraints for how microeconomic models can be manipulated.

Finally, by looking at the current uses and discussions of the rationality assumption in microeconomics textbooks, we see that even though the assumption in principle is decoupled from the notion of self-interest in some standard microeconomic models, the textbooks do not discuss this fact. Instead, they only employ examples with self-interested people. Further,

when the models use utilities rather than income, the textbooks still provide examples where individuals maximise their monetary income without stating or discussing the plausibility of the assumption that people's utilities will increase as a function of their income. Thus, microeconomics textbooks present the rationality assumption in all standard microeconomic models as if it includes the behavioural rule that people act (narrowly) self-interested in economic situations. Further, the informal discussions of the rationality assumption in the textbooks indicate that the assumption is both descriptively true and normatively desirable to follow. Thus, the textbooks do not live up to the standards of Friedman's methodological instrumentalism, and even when the textbooks describe what they are doing as positive economics, they slip into the domain of normative economics. This highlights a methodological point in philosophy of economics that even if it - in principle - is possible to keep a methodological divide between positive and normative economics, this in principle distinction is not maintained in economic theory today; and so we should ask whether it is defensible or desirable to keep insisting on it.

Using the above analyses, I argued that one way in which microeconomics can become self-fulfilling is by promoting a social norm of (narrow) self-interest in economic situations via the dissemination of microeconomics textbooks and teaching practices. In order to do this, I used Bicchieri's [2006] definition of social norms. According to this definition, a social norm exists and is followed if sufficiently many people know that a behavioural rule applies in certain situations (condition 1), and have a conditional preference for following that rule in those situations (condition 2) if they expect that sufficiently many people will follow the rule in the situations (condition 2(a)) and if they expect that sufficiently many people expect them to follow the rule in the situations (condition 2(b)). By using this definition as a starting point, I showed how readers of microeconomics textbooks can come to satisfy conditions 1, 2, 2(a), and 2(b) for a behavioural rule of (narrow) self-interest in economic situations due to the ways the rationality assumption is presented in the textbooks. Further, I argued that teaching practices at universities are structured in a way that confirms students' empirical and normative expectations. Thus, conditions 2(a) and 2(b) will stay satisfied for the students

such that they will keep following the norm. Looking at their social network, the hub created by the microeconomics class may thus be sufficient to spread the social norm to the entire network, making self-interested behaviour the predominant way to act in economic situations.

I ended the second part of the thesis by presenting the results of three experiments conducted to test the empirical plausibility of the above argument. The first two experiments confirm that economists and non-economists act differently in an economic situation (the prisoner's dilemma) and that differences in behaviour among participants are caused by a social norm of narrow self-interest and a social norm of cooperation being activated. The experiments also show that the norms can be activated via an individualist and a collectivist terminology designed to mirror the stimuli used in microeconomics textbooks and using the two microeconomic concepts of *rationality* and *social optimality*, respectively. Thus, the experimental results support the argument that microeconomics textbooks can promote a social norm of (narrow) self-interest among their readers in economic situations. Finally, I presented the results of a simulation experiment conducted to test how a social norm of narrow self-interest can spread in a social network. To do this, we constructed a behavioural model based on the results from the first experiment and used a social network constructed from the interactions between 1000 university students. The simulation methods we used were based on the methods from evolutionary game theory, which are often used to model emergent collective behaviour. The simulation results show that the fraction of nodes in the network that act like people exposed to the individualist terminology can completely change the outcome of the simulation, driving every node to cooperate or defect in a prisoner's dilemma game. Thus, the results support the argument that microeconomics teaching practices can stabilise a social norm of (narrow) self-interest in a social network by showing that it is possible to stabilise a social norm of narrow self-interest in a social network if sufficiently many nodes in the network act similarly to the participants exposed to the individualist terminology.

Thus, I conclude that scientific descriptions can become self-fulfilling by promoting a social norm among the people they are disseminated to.

Further, I conclude that microeconomics is currently self-fulfilling because it promotes a social norm of (narrow) self-interest via the dissemination of its textbooks and teaching practices. In addition to contributing to the debate in philosophy of science on self-fulfilling sciences, these findings also contribute to the literature in social ontology by showing that social norms exist and guide behaviour in economic situations.

Outlook and further discussion

The argument that microeconomics is a self-fulfilling science - where its textbooks and teaching practices can promote a social norm of (narrow) self-interest - can have important social implications. Recall from subsection 6.3.2 that economic situations include situations involving consumers and consumption, choices under uncertainty, coordination and competition with other agents, and bargaining or a market setting. Thus, the types of situations - for which a social norm of (narrow) self-interest is being promoted - go well beyond what lay people might perceive as an economic situation of buying and selling goods.[1]

Further, microeconomic theory is increasingly being taught and used both within and outside academia. This means that the number of people who are familiar with microeconomics textbook theory is increasing and so the scope of the self-fulfilling effects is increasing as well. This happens in at least three ways.

First, the discipline of economics has expanded dramatically since the second world war. This has happened both with regard to the number of universities having an economics department, the number of faculty member being economists, and the number of students taking economics degrees or courses [Fourcade, 2006, pp.162-163]. Thus, economics as a university degree is rapidly becoming more common and increasing its number of students. Further, students in government, law, and sociology will typically also have

[1]Examples of economic situations that we may not perceive as classical market situations can for example be found in the newspaper column *Undercover Economist* in the London Financial Times. Examples are "How can I win back my girlfriend?" and "Can cheap wine be a winner at dinner?" [Morgan, 2012, pp.403-404].

an introductory course to microeconomics, providing them only with the textbook version of economics considered in this thesis.

Second, economics influences policy making and institutional design to a greater extent than other social sciences:

> In the United States, for example, the President has the Council of Economic Advisers; there is no corresponding council for any other social science, even though other disciplines are pertinent to such social problems as welfare, work, criminology, and global affairs [Ferraro et al., 2005, p.11].

This is also the case in the UK where Earle et al. [2016, p.15] report that "the economy" was mentioned zero times in the manifesto of the winning party in the UK up until the 1940s. By the end of the 1980s, the number had increased to around 15. By the first half of the 2010s the number had exponentially increased to around 60. Thus, lay people and politicians are exposed to economic theory via its increased role in politics and political discourse.

Finally, lay people are increasingly exposed to microeconomic theory through popular economic articles and books intended for a non-academic audience such as Landsburg [2007], Hartford [2008], Smith [2010], Harford [2010], Frank [2011], Levitt and Dubner [2014a], Levitt and Dubner [2014b], Harford [2015], Levitt and Dubner [2015], and Raworth [2017]. Thus, the influence of microeconomics goes beyond the mere education of economics students affecting both politicians and lay people, and it is therefore important to consider whether it is beneficial for our societies to promote a social norm of (narrow) self-interest.

As discussed in chapter 4, the arguments historically used to support the benefits of self-interested behaviour - both for society and the individual - do not show that self-interested behaviour by itself is in fact beneficial. Indeed, Mandeville's argument is based on a very specific - and not commonly accepted - notion of vice and virtues, and Smith's comments in the *Wealth of Nations* do not support the later interpretations and scope of the "invisible hand argument". Further, the neoclassical arguments that self-interested actions can reduce discrimination lack a basic understanding of what discrimination is and that market forces are equally (if not more) likely

to increase discrimination rather than reducing it. Finally, game theory gives us several examples of how free riding and non-cooperative behaviour can be damaging for a society. Given these considerations, we do have reasons to avoid promoting a social norm of (narrow) self-interest. Whether there are sufficient reasons to change the current self-fulfilling effects of microeconomics textbook theory is outside the scope of the thesis. However, if we want to do so, the argument in this thesis help us understand how it can be done. Thus, I end the thesis with a short discussion of two ways it is possible to mitigate the current effects of microeconomics textbooks and teaching practices. The first and more demanding approach is to change the rationality assumption in microeconomic theory. The second and less demanding approach is to change the curriculum of microeconomics courses.

Mitigating social implications by changing the theory

The first approach - to change the rationality assumption in microeconomic theory - can hinder the promotion of a social norm of (narrow) self-interest by removing the assumption from economic theory. Even though the rationality assumption need not imply a behavioural rule of (narrow) self-interest in all economic models, it does so in consumer theory, theory of the firm, competitive markets, and general equilibrium theory. Thus, changing the behavioural assumption in these models can help mitigating the effects the models have on the people they are taught to. Further, changing the rationality assumption in other areas of microeconomics - such as choice under uncertainty and game theory - will ensure that microeconomics textbooks are unable to present the behavioural assumption as implying a behavioural rule of (narrow) self-interest. Thus, for all standard models in microeconomics, changing the behavioural assumptions so that they no longer imply (or can be interpreted as implying) only self-interested actions will ensure that people do not see the models as evidence that conditions 2(a) and 2(b) are satisfied in real economic situations.

Since the 1950s, there have been several attempts to formulate an alternative theory of human behaviour in microeconomics. One of the first attempts was Herbert Simon's suggestion that people's choice heuristic is

bounded [Simon, 1955, 1972]. Simon suggested to substitute the *optimizing* part of the rationality assumption with the assumption that people want a *satisfactory* outcome - above a certain boundary - rather than the best possible outcome. This, according to Simon [1955, p.100], is a more realistic account of how people choose, as people lack the computational abilities and time to gather information in order to optimize every decision. Today, bounded rationality describes a range of different theories that aim to account for choice behaviour for people who lack the computational abilities and time to be full rational agents [Wheeler, 2020, Klaes et al., 2005, Gigerenzer and Selten, 2002]. In the 1970s and early 1980s, additional alternative theories of human choice behaviour were developed as the critique of the rationality assumption increased [Sen, 1977]. These theories - such as prospect theory [Tversky and Kahneman, 1979, 1991] and later regret theory [Loomes and Sugden, 1982, Bell, 1982, Fishburn, 2013] - incorporated real psychological findings into their account of human decision making [Wheeler, 2020]. The attempts to incorporate different psychological motivations have continued since then, with - for example - Rabin [1993] suggesting the inclusion of reciprocity, Fehr and Schmidt [1999] proposing a model of inequality aversion that can account for altruistic and benevolent preferences, and Bicchieri [2006, ch.3] presenting a model that can account for the effect of social norms.[2]

Despite many attempts to develop new theories of human decision making in microeconomics, despite technological advances making mathematical simplicity superfluous, and despite the appeal of using a more realistic theory, the rationality assumption remains the predominant theory of human behaviour in microeconomics [Levin and Milgrom, 2004, p.3].[3] Thus, even if it is possible to mitigate the promotion of a social norm by changing

[2]For additional historical accounts of the development in decision theory within economics, see also Giocoli [2003] and Fontaine [2012].

[3]One reason why this may be - as discussed in section 5.4 - is that the rationality assumption is entangled with the definition of economics as a science [Giocoli, 2003, p.3] and with the use of mathematics to model economic situations. Further, the rationality assumption is used in all microeconomic models which creates a coherence between the different models that would not otherwise be there. That this contributes to the persistence of the assumption can be seen in informal discussions of the rationality assumption in the textbooks, where one of the main arguments in defence of the assumption was that it made the model consistent with the other microeconomic models.

the rationality assumption in microeconomic models, this may not be an attainable solution. I therefore turn to another and less demanding approach to mitigate the effects.

Mitigating social implications by changing the curriculum

The second approach to mitigate the social implications of microeconomic theory is to keep the rationality assumption as a part of microeconomic theory but to change the curriculum and teaching practices in microeconomics courses such that they no longer promote a social norm of (narrow) self-interest. This can be done by making sure that the textbooks and teaching practices do not give the impression that conditions 1 through 2(b) are satisfied for the norm in economic situations.

There are already a couple of textbooks available that do not promote a social norm of (narrow) self-interest in economic situations.[4] *Microeconomics in context* by Goodwin et al. [2014] is one such example. The first part of the book introduces the context of microeconomic analysis, including income inequality, taxes, differences in salaries, carbon dioxide emission, and the three spheres of economic activity; households, the public sector, and businesses. The second part presents a critical discussion of the market with supply and demand, price elasticity, welfare analysis, and international trade. In the third part, economic behaviour, consumption, and the labour market is discussed. For economic behaviour, the rationality assumption from neoclassical economics is clearly stated as such, and the textbook goes on to describe new approaches to modelling human behaviour. The final two parts of the book return to questions of inequality, taxes, the environment, and markets. Nowhere in the book is the rationality assumption endorsed as the appropriate or right way to act. Because of this, reading this textbook is unlikely to promote a social norm of (narrow) self-interest among its readers.

If university teachers are unable to change the curriculum in microeconomics courses to a textbook that does not promote a social norm of (narrow) self-interest, a minimally demanding way to avoid promoting this norm is

[4] See for example Goodwin et al. [2014], Komlos [2015], and Fischer et al. [2017].

to change the teaching practices of the courses. As mentioned in chapter 6, economics teaching practices generally consists of reading economics textbooks, going to lectures, attending class teaching, and solving problem sets. It is through the lectures, class teaching, and problem sets that microeconomics students are likely to have their empirical and normative expectations confirmed, so that they will keep following the norm of (narrow) self-interest. If this happens for sufficiently many economics students, then we know from the contagion literature (see subsection 6.2.3) and the simulation experiment described in section 7.4 that their interactions with other people in a social network can take the entire network over a tipping point such that everyone in the network will start following the norm. In order to avoid this, microeconomics teachers can - in the lectures, classes, and problem sets - spend time explaining that the assumption does not give an accurate description of human behaviour, nor a normatively desirable one. Further, the teachers can ask students to discuss and consider the assumption - and where it might fail - in the classes and problem sets. By making time for critical discussion of the assumption, students can observe that their peers do not act (narrowly) self-interested in economic situations (condition 2(a)), nor do they expect others to do so (condition 2(b)). In this case, students who are inclined to follow the norm will stop doing so, and thus the self-fulfilling effects of microeconomics will be mitigated.[5]

Some final remarks

Finally, the self-fulfilling effects of microeconomic theory can have social implication reaching beyond economics students due to its increased presence in political discourse and popular books. If we want to mitigate the social implications of microeconomics, it is thus relevant to consider the responsibilities of politicians, journalists, and authors of popular economics books. This does not mean that politicians should stop speaking about economics (far from it) but rather that they will have a responsibility to state when their arguments rest on assumptions that are not satisfied in the real

[5]For a thorough account of the different tools that can be used to change a social norm, see Bicchieri [2016, ch.3, 4, and 5].

world - such as the argument that reducing taxes will automatically make people work more. In the same fashion, journalists and authors will have a responsibility not to use positively laden words to present, explain, and use the rationality assumption in their analyses. By making these changes to how economic theory is presented and discussed, we can avoid promoting a social norm of (narrow) self-interest among the general public in economics situations.

Looking at the wider aim of the thesis, I used a case study of a particular self-fulfilling science in order to consider a general mechanism through which scientific descriptions can become self-fulfilling. Thus, the findings in the thesis can be used to inform other cases of self-fulfilling science by providing one explanation for how they can occur. If other sciences also become self-fulfilling by promoting social norms, then this lends support to a more general point about the power of language formulated - for example - by Julia Penelope [1990, p.213]:

> Language is power, in ways more literal than most people think. When we speak, we exercise the power of language to transform reality. Why don't more of us realize the connection between language and power?

Language is not always neutral, and when we create and use value-laden descriptions in science, we may change the world by promoting social norms that influence people's behaviour.

Bibliography

Robert P Abelson. Psychological status of the script concept. *American psychologist*, 36(7):715, 1981.

Ali Ahmed. Can education affect pro-social behavior? cops, economists and humanists in social dilemmas. *International Journal of Social Economics*, 35 (4):298–307, 2008.

George A Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.

Andrew Altman. Discrimination. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

Craig A Anderson. Imagination and expectation: The effect of imagining behavioral scripts on personal influences. *Journal of personality and social psychology*, 45(2):293, 1983.

Elizabeth Anderson. Uses of value judgments in science: A general argument, with lessons from a case study of feminist research on divorce. *Hypatia*, 19 (1):1–24, 2004.

Kenneth J Arrow and Gerard Debreu. Existence of an equilibrium for a competitive economy. *Econometrica: Journal of the Econometric Society*, pages 265–290, 1954.

K.J. Arrow and F. Hahn. *General Competitive Analysis*. Advanced textbooks in economics. North-Holland, edition: 1980, 1971. ISBN 9780444854971. URL https://books.google.dk/books?id=cKy2AAAAIAAJ.

Solomon E Asch. Effects of group pressure upon the modification and distortion of judgments. *Organizational influence processes*, pages 295–303, 1951.

Robert Aumann. Nash equilibria are not self-enforcing. *Economic Decision Making: Games, Econometrics and Optimisation*, pages 201–206, 1990.

John Langshaw Austin. *How to do things with words*. Oxford university press, 1975.

Robert Axelrod. An evolutionary approach to norms. *American political science review*, 80(4):1095–1111, 1986.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

John A Bargh and Tanya L Chartrand. The mind in the middle. *Handbook of research methods in social and personality psychology*, 2:253–285, 2000.

Frederic Charles Bartlett. *Remembering: A study in experimental and social psychology*. Cambridge University Press, 1995, 1932.

William M Baum, Brian Paciotti, Peter Richerson, Mark Lubell, and Richard McElreath. Cooperation due to cultural norms, not individual reputation. *Behavioural processes*, 91(1):90–93, 2012.

Yoram Bauman and Elaina Rose. Selection or indoctrination: Why do economics students donate less than the rest? *Journal of Economic Behavior & Organization*, 79(3):318–327, 2011.

Roy F Baumeister and Lauren E Brewer. Believing versus disbelieving in free will: Correlates and consequences. *Social and Personality Psychology Compass*, 6(10):736–745, 2012.

Anabel Belaus, Cecilia Reyna, and Esteban Freidin. Testing the effect of cooperative/competitive priming on the prisoner's dilemma. a replication study. *PloS one*, 13(12):e0209263, 2018.

David E Bell. Regret in decision making under uncertainty. *Operations research*, 30(5):961–981, 1982.

Carsten Bergenholtz and Jacob Busch. Self-fulfillment of social science theories: Cooling the fire. *Philosophy of the Social Sciences*, 46(1):24–43, 2016.

Adam J Berinsky, Gregory A Huber, and Gabriel S Lenz. Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. *Political analysis*, 20(3):351–368, 2012.

Ana Maria Bianchi. Are brazilian economists different? *Revista Brasileira de Economia*, 52(3):427–440, 1998.

Cristina Bicchieri. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, 2006.

Cristina Bicchieri. *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press, 2016.

Cristina Bicchieri and Peter McNally. Shrieking sirens: Schemata, scripts, and social norms. how change occurs. *Social Philosophy and Policy*, 35(1): 23–53, 2018.

Cristina Bicchieri and Erte Xiao. Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22(2):191–208, 2009.

Cristina Bicchieri, Ryan Muldoon, and Alessandro Sontuoso. Social norms. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2018 edition, 2018.

Thomas Blass. The milgram paradigm after 35 years: Some things we now know about obedience to authority 1. *Journal of applied social psychology*, 29 (5):955–978, 1999.

Lawrence E. Blume. The Statistical Mechanics of Strategic Interaction. *Games and Economic Behavior*, 5(3):387–424, jul 1993. ISSN 0899-8256. doi: 10.1006/GAME.1993.1023. URL https://www.sciencedirect.com/science/article/pii/S0899825683710237.

Stefan Bode, Anna Hanxi He, Chun Siong Soon, Robert Trampel, Robert Turner, and John-Dylan Haynes. Tracking the unconscious generation of free decisions using uitra-high field fmri. *PloS one*, 6(6):e21612, 2011.

Niels Bohr. *The structure of the atom, Nobel prize lecture*. Elsevier publishing company, Amsterdam 1965, 1922.

Niels Bohr. *The quantum postulate and the recent development of atomic theory*. Nature Publishing Group, 1928.

Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415): 295–298, 2012.

Shane Bonetti. Experimental economics and deception. *Journal of Economic Psychology*, 19(3):377–395, 1998.

Emma J Boyland and Jason CG Halford. Television advertising and branding. effects on eating behaviour and food preferences in children. *Appetite*, 62: 236–241, 2013.

Michael Emmett Brady. Adam smith was consistent in both the theory of moral sentiments and the wealth of nations on the role of the concept of self interest: Das utilitarian economist view is the problem. *Available at SSRN 3156013*, 2018.

William F Brewer and James C Treyens. Role of schemata in memory for places. *Cognitive psychology*, 13(2):207–230, 1981.

Nicolas Brisset. Economics is not always performative: some limits for performativity. *Journal of Economic Methodology*, 23(2):160–184, 2016.

Jeannette Brosig, Timo Heinrich, Thomas Riechmann, Ronnie Schöb, and Joachim Weimann. Laying off or not? the influence of framing and economics education. *International Review of Economics Education*, 9(1): 44–55, 2010.

Kamilla Haworth Buchter, Bjarke Mønsted, and Sune Lehmann. Self-interested behaviour as a social norm. *arXiv e-prints*, art. arXiv:2008.01884, August 2020.

Roger C Buck. Reflexive predictions. *Philosophy of Science*, 30(4):359–369, 1963.

Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.

Jerry M Burger. Replicating milgram: Would people still obey today? *American Psychologist*, 64(1):1, 2009.

Rick Busselle. Schema theory and mental models. *The international encyclopedia of media effects*, pages 1–8, 2017.

Donn Byrne and RA Baron. *Social psychology: understanding human interaction*. Boston; Toronto: Allyn and Bacon, 3rd edition. First published: 1977, 1981.

Bruce Caldwell. The chicago school, hayek, and neoliberalism. *Building Chicago Economics: New Perspectives on the History of America's Most Powerful Economics Program*, pages 301–34, 2011.

Michel Callon. Four models for the dynamics of science. In *Science and the Quest for Reality*, pages 249–292. Springer, 1995.

Michel Callon. Introduction: the embeddedness of economic markets in economics. *The sociological review*, 46:1–57, 1998.

Tom Campbell. *Adam Smith's science of morals*. Routledge, 2014.

Nancy Cantor, Walter Mischel, and Judith C Schwartz. A prototype analysis of psychological situations. *Cognitive psychology*, 14(1):45–77, 1982.

Alexander W Cappelen, Knut Nygaard, Erik Ø Sørensen, and Bertil Tungodden. Social preferences in the lab: A comparison of students and a representative population. *The Scandinavian Journal of Economics*, 117(4): 1306–1326, 2015.

John R Carter and Michael D Irons. Are economists different, and if so, why? *Journal of Economic Perspectives*, 5(2):171–177, 1991.

Nancy Cartwright. Mill and menger: Ideal elements and stable tendencies. *Poznan Studies in the Philosophy of the Sciences and the Humanities*, 38:171–188, 1994.

Damon Centola. The spread of behavior in an online social network experiment. *science*, 329(5996):1194–1197, 2010.

Larry Christensen. Deception in psychological research: When is its use justified? *Personality and Social Psychology Bulletin*, 14(4):664–675, 1988.

Robert B Cialdini and Noah J Goldstein. Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55:591–621, 2004.

Robert B Cialdini, Raymond R Reno, and Carl A Kallgren. A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology*, 58(6):1015, 1990.

Giam Pietro Cipriani, Diego Lubian, and Angelo Zago. Natural born economists? *Journal of Economic Psychology*, 30(3):455–468, 2009.

Franck Cochoy. Another discipline for the market economy: marketing as a performative knowledge and know-how for capitalism. *The Sociological Review*, 46(1_suppl):194–221, 1998.

Alain Cohn and Michel André Maréchal. Priming in economics. *Current Opinion in Psychology*, 12:17–21, 2016.

Alan M Colins and Eizabeth F Loftus. A spreading-activation theory of semantic memory. *Psychological Review*, 82:407–428, 1975.

Allan M Collins, M Ross Quillian, et al. Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, 8(2):240–247, 1969.

W J Conover and Ronald L Iman. On multiple-comparisons procedures. *Los Alamos Sci. Lab. Tech. Rep. LA-7677-MS*, pages 1–14, 1979.

Richard Cookson. Framing effects in public goods experiments. *Experimental Economics*, 3(1):55–79, 2000.

David J Cooper. A note on deception in economic experiments. *Journal of Wine Economics*, 9(2):111–114, 2014.

Rachel Croson and Melanie Marks. Identifiability of individual contributions in a threshold public goods experiment. *Journal of Mathematical Psychology*, 42(2-3):167–190, 1998.

Dean Curran. From performativity to representation as intervention: Rethinking the 2008 financial crisis and the recent history of social science. *Journal for the Theory of Social Behaviour*, 48(4):492–510, 2018.

Micael Dahlen, Helge Thorbjørnsen, Jonas Colliander, Sara Rosengren, Alice Gemvik, and Christian Thorwid. The effects of communicating passion in advertising. *Journal of Advertising Research*, 60(1):3–11, 2020. ISSN 0021-8499. doi: 10.2501/JAR-2019-040. URL http://www.journalofadvertisingresearch.com/content/60/1/3.

Emmanuelle de Champs. Happiness and interests in politics: A late-enlightenment debate. *Happiness and Utility: Essays Presented to Frederick Rosen*, page 20, 2019.

Paul-Henri d'Holbach. *System of Nature*. New York: Burt Franklin, Translated by HD Robinson, 1970, 1770.

Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 135–143, 2018.

Wen-Bo Du, Hao-Ran Zheng, and Mao-Bin Hu. Evolutionary prisoner's dilemma game on weighted scale-free networks. *Physica A: Statistical Mechanics and its Applications*, 387(14):3796–3800, 2008.

Martin Dufwenberg, Simon Gächter, and Heike Hennig-Schmidt. The framing of games and the psychology of play. *Games and Economic Behavior*, 73(2): 459–478, 2011.

Margaret Carlisle Duncan. Sports photographs and sexual difference: Images of women and men in the 1984 and 1988 olympic games. *Sociology of sport journal*, 7(1):22–43, 1990.

Joe Earle, Cahal Moran, and Zach Ward-Perkins. *The econocracy*. Manchester University Press, 2016.

Asia A Eaton, Suzanna M Rose, Camille Interligi, Katherine Fernandez, and Maureen McHugh. Gender and ethnicity in dating, hanging out, and hooking up: Sexual scripts among hispanic and white young adults. *The Journal of Sex Research*, 53(7):788–804, 2016.

Tore Ellingsen, Magnus Johannesson, Johanna Mollerstrom, and Sara Munkhammar. Social framing effects: Preferences or beliefs? *Games and Economic Behavior*, 76(1):117–130, 2012.

Christoph Engel and David G Rand. What does "clean" really mean? the implicit framing of decontextualized experiments. *Economics Letters*, 122 (3):386–389, 2014.

Jerry Evensky. *Adam Smith's moral philosophy: a historical and contemporary perspective on markets, law, ethics, and culture*. Cambridge University Press, 2005.

Marco Faravelli. How context matters: A survey based experiment on distributive justice. *Journal of Public Economics*, 91(7-8):1399–1422, 2007.

Ernst Fehr and Klaus M Schmidt. A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868, 1999.

Ernst Fehr and Ivo Schurtenberger. Normative foundations of human cooperation. *Nature human behaviour*, 2(7):458, 2018.

Teppo Felin and Nicolai J Foss. Performativity of theory, arbitrary conventions, and possible worlds: A reality check. *Organization Science*, 20(3):676–678, 2009a.

Teppo Felin and Nicolai J Foss. Social reality, the boundaries of self-fulfilling prophecy, and economics. *Organization Science*, 20(3):654–668, 2009b.

Fabrizio Ferraro, Jeffrey Pfeffer, and Robert I Sutton. Economics language and assumptions: How theories can become self-fulfilling. *Academy of Management review*, 30(1):8–24, 2005.

Fabrizio Ferraro, Jeffrey Pfeffer, and Robert I Sutton. How and why theories matter: A comment on felin and foss (2009). *Organization Science*, 20(3): 669–675, 2009.

Annamaria Fiore. Experimental economics: some methodological notes. *MPRA*, 2009.

Liliann Fischer, Joe Hasell, J Christopher Proctor, David Uwakwe, Zach Ward Perkins, and Catriona Watson. *Rethinking economics: An introduction to pluralist economics*. Routledge, 2017.

Peter Fischer, Joachim I Krueger, Tobias Greitemeyer, Claudia Vogrincic, Andreas Kastenmüller, Dieter Frey, Moritz Heene, Magdalena Wicher, and Martina Kainbacher. The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological bulletin*, 137(4):517, 2011.

Martin Fishbein and Icek Ajzen. *Predicting and changing behavior: The reasoned action approach*. Taylor & Francis, 2011.

Peter C Fishburn. *The foundations of expected utility*, volume 31. Springer Science & Business Media, 2013.

Athol Fitzgibbons and Athol Fitzgibbons. *Adam Smith's system of liberty, wealth, and virtue: The moral and political foundations of the wealth of nations*. Oxford University Press, 1995.

M Flood, M Dresher, A Tucker, and F Device. Prisoner's dilemma: Game theory. In *Experimental Economics*. Rand Corporation, 1950.

Philippe Fontaine. Beyond altruism? economics and the minimization of unselfish behavior, 1975–93. *History of Political Economy*, 44(2):195–233, 2012.

Pierre Force. *Self-interest before Adam Smith: A genealogy of economic science*, volume 68. Cambridge University Press, 2003.

Marion Fourcade. The construction of a global profession: The transnationalization of economics. *American journal of sociology*, 112(1):145–194, 2006.

Björn Frank and Günther G Schulze. Does economics make citizens corrupt? *Journal of economic behavior & organization*, 43(1):101–113, 2000.

Robert H Frank. *The economic naturalist: Why economics explains almost everything*. Random House, 2011.

Robert H Frank, Thomas Gilovich, and Dennis T Regan. Does studying economics inhibit cooperation? *Journal of economic perspectives*, 7(2):159–171, 1993.

Robert H Frank, Thomas D Gilovich, and Dennis T Regan. Do economists make bad citizens? *Journal of Economic Perspectives*, 10(1):187–192, 1996.

Bruno S Frey and Stephan Meier. Are political economists selfish and indoctrinated? evidence from a natural experiment. *Economic Inquiry*, 41 (3):448–462, 2003.

Bruno S Frey, Werner W Pommerehne, and Beat Gygi. Economics indoctrination or selection? some empirical results. *The Journal of Economic Education*, 24(3):271–281, 1993.

Miranda Fricker. *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press, 2007.

Itzhak Fried, Roy Mukamel, and Gabriel Kreiman. Internally generated preactivation of single neurons in human medial frontal cortex predicts volition. *Neuron*, 69(3):548–562, 2011.

Milton Friedman. The methodology of positive economics. *Essays in positive economics*, 3(3):145–178, 1953.

Milton Friedman. *Capitalism and freedom*. University of Chicago press, 2009, 1962.

Milton Friedman. The need for futures markets in currencies. *Cato J.*, 31:635, 1971.

Roman Frigg and Stephan Hartmann. Models in science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 edition, 2018.

Francis Fukuyama. *The great disruption*. Profile Books, 2017.

Adrian Furnham and Hua Chu Boo. A literature review of the anchoring effect. *The journal of socio-economics*, 40(1):35–42, 2011.

David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

Neil Gandal, Sonia Roccas, Lilach Sagiv, and Amy Wrzesniewski. Personal value priorities of economists. *Human Relations*, 58(10):1227–1252, 2005.

MM Gergen and KJ Gergen. *Social psychology*. New York; Harcourt Brace Jovanovich, 1981.

Philipp Gerlach. The games economists play: Why economics students behave more selfishly than other students. *PloS one*, 12(9):e0183814, 2017.

Philipp Gerlach and Bastian Jaeger. Another frame, another game? explaining framing effects in economic games. *Norms, Actions, Games*, 2016.

Allan Gibbard and Hal R Varian. Economic models. *The Journal of Philosophy*, 75(11):664–677, 1978.

Robert Gibbons et al. *A primer in game theory*. Harvester Wheatsheaf New York, 1992.

Gerd Gigerenzer and Reinhard Selten. *Bounded rationality: The adaptive toolbox*. MIT press, 2002.

Nicola Giocoli. *Modeling rational agents: From interwar economics to early modern game theory*. Edward Elgar Publishing, 2003.

Nathan Glazer. Interests and passions. *The Public Interest*, 81:17, 1985.

Neva Goodwin, Julie Nelson, Jonathan Harris, Mariano Torras, and Brian Roach. *Microeconomics in context, third edition*. ME Sharpe, 2014.

Carlos Gracia-Lázaro, José A Cuesta, Angel Sánchez, and Yamir Moreno. Human behavior in Prisoner's Dilemma experiments suppresses network reciprocity. *Scientific reports*, 2:325, 2012. ISSN 2045-2322. doi: 10.1038/srep00325.

Carlos Gracia-Lázaro, José A Cuesta, Angel Sánchez, and Yamir Moreno. Human behavior in prisoner's dilemma experiments suppresses network reciprocity. *Scientific reports*, 2(1):1–4, 2012.

Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

Mark Granovetter. *Getting a job: A study of contacts and careers*. University of Chicago press, 2018.

Ruth W Grant. Passions and interests revisited: the psychological foundations of economics and politics. *Public Choice*, 137(3-4):451–461, 2008.

Carlos Grilo and Luís Correia. The influence of the update dynamics on the evolution of cooperation. *International Journal of Computational Intelligence Systems*, 2(2):104–114, 2009.

Alan E Gross and India Fleming. Twenty years of deception in social psychology. *Personality and Social Psychology Bulletin*, 8(3):402–408, 1982.

Jelena Grujić, Constanza Fosco, Lourdes Araujo, José A. Cuesta, and Angel Sánchez. Social Experiments in the Mesoscale: Humans Playing a Spatial Prisoner's Dilemma. *PLoS ONE*, 5(11):e13749, nov 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0013749. URL https://dx.plos.org/10.1371/journal.pone.0013749.

Adolf Grünbaum. Historical determinism, social activism, and predictions in the social sciences. *The British Journal for the Philosophy of Science*, 7(27): 236–240, 1956.

Francesco Guala. Performativity rationalized. In *Enacting Dismal Science*, pages 29–52. Springer, 2016a.

Francesco Guala. *Understanding institutions: The science and philosophy of living together*. Princeton University Press, 2016b.

Stephen Gudeman. *Anthropology and economy*. Cambridge University Press, 2016.

Andrés M Guiot-Isaac. Latin america through the eyes of albert o. hirschman. *Ideas in the History of Economic Development: The Case of Peripheral Countries*, 2019.

Mazen Maurice Guirguis. *A script theory of intentional content*. PhD thesis, University of British Columbia, 2003.

Ian Hacking. The looping effects of human kinds. In Dan Sperber, David Premack, and Ann James Premack, editors, *Causal cognition: A multidisciplinary debate.*, page 351–394. Clarendon Press/Oxford University Press, New York, NY, US, 1995.

Bert Hamminga and Neil De Marchi. *Idealization VI: Idealization in economics*. Rodopi, 1994.

James A Hampton. *Categories, prototypes and exemplars*. Routledge: Abington, UK, 2016.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Benjamin V Hanrahan, Jeffrey P Bigham, and Chris Callison-Burch. Worker demographics and earnings on amazon mechanical turk: An exploratory analysis. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.

Tim Harford. *The undercover economist*. Hachette UK, 2010.

Tim Harford. *The Undercover Economist Strikes Back: How to Run, Or Ruin, an Economy*. Penguin, 2015.

PD Harms and Justin A DeSimone. Caution! mturk workers ahead—fines doubled. *Industrial and Organizational Psychology*, 8(2):183–190, 2015.

John C Harsanyi. A new theory of equilibrium selection for games with complete information. *Games and Economic Behavior*, 8(1):91–122, 1995.

John C Harsanyi, Reinhard Selten, et al. A general theory of equilibrium selection in games. *MIT Press Books*, 1, 1988.

Tom Hartford. *The logic of life*. New York: Little Brown, 2008.

Justus Haucap and Tobias Just. Not guilty? another look at the nature and nurture of economics students. *European Journal of Law and Economics*, 29 (2):239–254, 2010.

Justus Haucap and Andrea Müller. Why are economists so different? nature, nurture and gender effects in a simple trust game. *Nature, Nurture and Gender Effects in a Simple Trust Game (March 1, 2014)*, 2014.

David J Hauser and Norbert Schwarz. Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, 48(1):400–407, 2016.

Scott A Hawkins, Stephen J Hoch, and Joan Meyers-Levy. Low-involvement learning: Repetition and coherence in familiarity and belief. *Journal of Consumer Psychology*, 11(1):1–11, 2001.

Friedrich August Hayek. *The road to serfdom: Text and documents: The definitive edition*. Routledge, 2014, 1944.

Friedrich August Hayek. Lecture on a master mind. In *Proceedings of the British Academy*, volume 52, pages 125–141, 1966.

Friedrich August Hayek. The results of human action but not of human design. *Studies in philosophy, politics and economics*, 96, 1967.

Friedrich August Hayek. *Law, legislation and liberty, volume 2: The mirage of social justice*, volume 2. University of Chicago Press, 2012, 1976.

Jeffery Herbener. An integration of the wealth of nations and the theory of moral sentiments. *The Journal of Libertarian Studies*, 8(2):275–288, 1987.

Miles Hewstone. Revision and change of stereotypic beliefs: In search of the elusive subtyping model. *European review of social psychology*, 5(1):69–109, 1994.

Albert O Hirschman. *The passions and the interests: Political arguments for capitalism before its triumph*. Greenwood Publishing Group, 1997.

Li-Ming Ho, Christian Schafferer, Jie-Min Lee, Chun-Yuan Yeh, and Chi-Jung Hsieh. Raising cigarette excise tax to reduce consumption in low-and middle-income countries of the asia-pacific region: a simulation of the anticipated health and taxation revenues impacts. *BMC public health*, 18(1): 1187, 2018.

Thomas Hobbes. *Leviathan*. A&C Black, 2006, 1651.

Astri Drange Hole. How do economists differ from others in distributive situations? *Nordic Journal of Political Economy*, 38, 2013.

Petter Holme, Ala Trusina, Beom Jun Kim, and Petter Minnhagen. Prisoners' dilemma in real-world acquaintance networks: Spikes and quasiequilibria induced by the interplay between structure and dynamics. *Physical Review E*, 68(3):030901, 2003.

Thomas A Horne. Envy and commercial society: Mandeville and smith on" private vices, public benefits". *Political Theory*, 9(4):551–569, 1981.

Yung-An Hu and Day-Yang Liu. Altruism versus egoism in human behavior of mixed motives: An experimental study. *American Journal of Economics and Sociology*, 62(4):677–705, 2003.

David Hume. *A treatise of human nature*. Courier Corporation, 2003, 1738.

Rosalind Hursthouse and Glen Pettigrove. Virtue ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2018 edition, 2018.

John Ifcher and Homa Zarghamee. The rapid evolution of homo economicus: Brief exposure to neoclassical assumptions increases self-interested behavior. *Journal of Behavioral and Experimental Economics*, 75:55–65, 2018.

Maki Inoue-Choi, Scott P Kelly, Kelvin Choi, and Neal D Freedman. *Lifetime trajectories of cigarette smoking and cancer mortality among older adults in a large cohort in the United States*. AACR, 2018.

Struan Jacobs. Spontaneous order: Michael polanyi and friedrich hayek. *Critical Review of International Social and Political Philosophy*, 3(4):49–67, 2000.

Annamarie Jagose. *Queer theory: An introduction*. NYU Press, 1996.

Sarwat Jahan, Ahmed Saber Mahmud, and Chris Papageorgiou. What is keynesian economics? *International Monetary Fund*, 51(3), 2014.

Tammy James, Lewis Soroka, and John G Benjafield. Are economists rational, or just different? *Social Behavior and Personality: an international journal*, 29 (4):359–364, 2001.

Richard C Jeffrey. *The logic of decision*. University of Chicago Press, 1990.

Geoffrey Alexander Jehle and Philip J. Reny. *Advanced microeconomic theory*. Pearson Education India, 2011.

William Stanley Jevons. *The theory of political economy, Augustus M.* Kelley, New York, 1871.

Herman Judith et al. Trauma and recovery. *London: Pandora*, 1992.

Aaron C Kay and Lee Ross. The perceptual push: The interplay of implicit cues and explicit situational construals on behavioral intentions in the prisoner's dilemma. *Journal of Experimental Social Psychology*, 39(6):634–643, 2003.

Aaron C Kay, S Christian Wheeler, John A Bargh, and Lee Ross. Material priming: The influence of mundane physical objects on situational construal and competitive behavioral choice. *Organizational behavior and human decision processes*, 95(1):83–96, 2004.

Jeremy Kees, Christopher Berry, Scot Burton, and Kim Sheehan. An analysis of data quality: Professional panels, student subject pools, and amazon's mechanical turk. *Journal of Advertising*, 46(1):141–155, 2017.

Gavin Kennedy. Adam smith and the invisible hand: From metaphor to myth. *Econ Journal Watch*, 6(2):239, 2009.

Claudia Keser. Voluntary contributions to a public good when partial contribution is a dominant strategy. *Economics Letters*, 50(3):359–366, 1996.

John Maynard Keynes. The general theory of employment. *The quarterly journal of economics*, 51(2):209–223, 1936.

Sung Ho Kim. Max weber. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2019 edition, 2019.

Matthias Klaes, Esther-Mirjam Sent, et al. A conceptual history of the emergence of bounded rationality. *History of political economy*, 37(1):27–59, 2005.

Arjo Klamer and David Colander. *The making of an economist*. Routledge, 2019 edition, 1990.

Richard Klein, Kate Ratliff, Michelangelo Vianello, Reginald Adams Jr, Stěpán Bahník, Michael Bernstein, Konrad Bocian, Mark Brandt, Beach Brooks, Claudia Brumbaugh, et al. Data from investigating variation in replicability: A "many labs" replication project. *Journal of Open Psychology Data*, 2(1), 2014.

Marc Knez and Colin Camerer. Increasing cooperation in prisoner's dilemmas by establishing a precedent of efficiency in coordination games. *Organizational Behavior and Human Decision Processes*, 82(2):194–216, 2000.

Frank H Knight. *Risk, uncertainty and profit*. Courier Corporation, 2012, 1921.

John Komlos. *What every economics student needs to know and doesn't get in the usual principles text*. Routledge, 2015.

Matthew Kopec. A more fulfilling (and frustrating) take on reflexive predictions. *Philosophy of science*, 78(5):1249–1259, 2011.

Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

Max M Krasnow, Rhea M Howard, and Adar B Eisenbruch. The importance of being honest? evidence that deception may not pollute social science subject pools after all. *Behavior research methods*, pages 1–14, 2019.

Harold William Kuhn and Albert William Tucker. *Contributions to the Theory of Games*, volume 2. Princeton University Press, 1953.

Thomas S Kuhn. Objectivity, value judgment, and theory choice. *Arguing about science*, pages 74–86, 1977.

David N Laband and Richard O Beil. Are economists more selfish than other'social'scientists? *Public Choice*, 100(1-2):85–101, 1999.

Hugh Lacey. *Is science value free?: Values and scientific understanding*. Psychology Press, 2004.

Maurice Lagueux. *Was Keynes a liberal and an individualist? Or Keynes reader of Mandeville*. L'Harmattan, 1998.

Steven E Landsburg. *The Armchair Economist (revised and updated May 2012): Economics & Everyday Life*. Simon and Schuster, 2007.

Alessandro Lanteri. (why) do selfish people self-select in economics? *Erasmus Journal for Philosophy and Economics*, 1(1):1–23, 2008.

Bibb Latane and John M Darley. Group inhibition of bystander intervention in emergencies. *Journal of personality and social psychology*, 10(3):215, 1968.

Bruno Latour. *Science in action: How to follow scientists and engineers through society*. Harvard university press, 1987.

Robert K Lech, Onur Güntürkün, and Boris Suchan. An interplay of fusiform gyrus and hippocampus enables prototype-and exemplar-based category learning. *Behavioural Brain Research*, 311:239–246, 2016.

Yun Shin Lee, Yong Won Seo, and Enno Siemsen. Running behavioral operations experiments using amazon's mechanical turk. *Production and Operations Management*, 27(5):973–989, 2018.

Abba P Lerner. Stagflation–its cause and cure. *Challenge*, 20(4):14–19, 1977.

Jonathan Levin and Paul Milgrom. Introduction to choice theory. *Available from internet: http://web. stanford. edu/˜ jdlevin/Econ*, 20202, 2004.

Steven D Levitt and Stephen J Dubner. *Freakonomics*. B DE BOOKS, 2014a.

Steven D Levitt and Stephen J Dubner. *Think like a freak*. Harper Audio, 2014b.

Steven D Levitt and Stephen J Dubner. *When to Rob a Bank: A Rogue Economist's Guide to the World*. Penguin UK, 2015.

Varda Liberman, Steven M Samuels, and Lee Ross. The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Personality and social psychology bulletin*, 30(9):1175–1185, 2004.

Benjamin Libet, Curtis A Gleason, Elwood W Wright, and Dennis K Pearl. Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). In *Neurophysiology of Consciousness*, pages 249–268. Springer, 1993.

HC Lindgren and JH Harvey. *An introduction to social psychology, 3rd edition*. St. Louis; Mosby, 1981.

Kasper Lippert-Rasmussen. The badness of discrimination. *Ethical Theory and Moral Practice*, 9(2):167–185, 2006.

Christian List and Philip Pettit. *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press, 2011.

Helen E Longino. *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press, 1990.

Helen E Longino. In search of feminist epistemology. *The monist*, 77(4): 472–485, 1994.

Graham Loomes and Robert Sugden. Regret theory: An alternative theory of rational choice under uncertainty. *The economic journal*, 92(368):805–824, 1982.

João Carlos Lopes, João Carlos Graça, and Rita Gomes Correia. Effects of economic education on social and political values, beliefs and attitudes: Results from a survey in portugal. *Procedia Economics and Finance*, 30: 468–475, 2015.

Charles Lowe. The significance of self-fulfilling science. *Philosophy of the Social Sciences*, 48(4):343–363, 2018.

Alexander L Macfie. Adam smith's moral sentiments as foundation for his wealth of nations. *Oxford Economic Papers*, 11(3):209–228, 1959.

Niccolò Machiavelli. *The prince*. Hackett Publishing, 2008, 1532.

Donald MacKenzie. *An engine, not a camera: How financial models shape markets*. Mit Press, 2008.

Uskali Mäki, editor. *The Methodology of Positive Economics: Reflections on the Milton Friedman Legacy*. Cambridge University Press, United Kingdom, 2009.

Uskali Mäki. Realism and antirealism about economics. *Philosophy of economics*, 13:3–24, 2012.

Uskali Mäki. Performativity: Saving austin from mackenzie. In *EPSA11 perspectives and foundational problems in philosophy of science*, pages 443–453. Springer, 2013.

Thomas Robert Malthus. *An essay on the principle of population, as it affects the future improvement of society*. J. Johnson, London, Electronic Scholarly Publishing Project 1998, 1798.

Bernard Mandeville. *The fable of the bees*. Liberty Fund, 1988, 1714.

Jean Matter Mandler. *Stories, scripts, and scenes: Aspects of schema theory*. Psychology Press, 2014.

Harvey C Mansfield. Self-interest rightly understood. *Political Theory*, 23(1): 48–66, 1995.

Gerald Marwell and Ruth E Ames. Economists free ride, does anyone else?: Experiments on the provision of public goods, iv. *Journal of public economics*, 15(3):295–310, 1981.

Karl Marx. Capital. a critique of political economy. volume i: Book one: The process of production of capital. *Moscow, RU: Progress Publishers*, 1887.

Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. *Microeconomic theory*, volume 1. Oxford university press New York, 1995.

Deirdre N McCloskey. *The bourgeois virtues: Ethics for an age of commerce*. University of Chicago Press, 2010.

W McElroy. The origin of religious tolerance: Voltaire's enlightening observations. *FREEMAN-NEW YORK-FOUNDATION FOR ECONOMIC EDUCATION-*, 48:347–350, 1998.

Lionel McKenzie. On equilibrium in graham's model of world trade and other competitive systems. *Econometrica: Journal of the Econometric Society*, pages 147–161, 1954.

Carl Menger. *Principles of Economics*. Ludwig von Mises Institute, Auburn, Alabama, 1976. Reprinted 2007, 1871.

Carl Menger. *Investigations into the Method of the Social Sciences*. Ludwig von Mises Institute, 1996, 1883.

Robert K Merton. Theself-fulfillingprophecy. *TheAntioch Review8 (2)*, pages 193–210, 1948.

Robert K Merton. The thomas theorem and the matthews effect. *Soc. F.*, 74: 379, 1995.

Robert King Merton. *Social theory and social structure*. Simon and Schuster, edt. Merton, Robert C, 1968, 1949.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182, 2011.

Stanley Milgram. Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4):371, 1963.

John Stuart Mill. Essay v: On the definition of political economy; and on the method of investigation proper to it. *Essays on some unsettled questions of political economy*, pages 86–114, 1836.

John Stuart Mill. *The Collected Works of John Stuart Mill, Volume II: The Principles of Political Economy with Some of Their Applications to Social Philosophy*. Online Library of Liberty, 1848.

Dale T Miller and Rebecca K Ratner. The disparity between the actual and assumed power of self-interest. *Journal of personality and social psychology*, 74(1):53, 1998.

H Laurence Miller Jr. On the" chicago school of economics". *Journal of Political Economy*, 70(1):64–69, 1962.

Karl Ove Moene. The moral sentiments of wealth of nations. *Adam Smith Review*, 6, 2011.

Daniel C Molden. Understanding priming effects in social psychology: What is "social priming" and how does it occur? *Social Cognition*, 32(Supplement): 1–11, 2014.

Andrew L Molinsky, Adam M Grant, and Joshua D Margolis. The bedside manner of homo economicus: How and why priming an economic schema reduces compassion. *Organizational Behavior and Human Decision Processes*, 119(1):27–37, 2012.

Bjarke Mønsted, Piotr Sapieżyński, Emilio Ferrara, and Sune Lehmann. Evidence of complex contagion of information in social media: An experiment using twitter bots. *PloS one*, 12(9):e0184148, 2017.

Leonidas Montes. The origins of das adam smith problem and our understanding of sympathy. *The Street Porter and the Philosopher: Conversations on Analytical Egalitarianism*, pages 158–78, 2008.

Mary S Morgan. *The world in the model: How economists work and think*. Cambridge University Press, 2012.

Ivan Moscati. *Measuring utility: From the marginal revolution to behavioral economics*. Oxford Studies in History of E, 2018.

Matthew Mulford, Jonathan Jackson, and Henrik Svedsäter. Encouraging cooperation: Revisiting solidarity and commitment effects in prisoner's dilemma games 1. *Journal of Applied Social Psychology*, 38(12):2964–2989, 2008.

Ernest Nagel. *The structure of science; problems in the logic of scientific explanation.* Harcourt, Brace & World, New York, 1961.

John Nash. Non-cooperative games. *Annals of mathematics*, pages 286–295, 1951.

David Newth and David Cornforth. Asynchronous spatial evolutionary games. *BioSystems*, 95(2):120–129, 2009.

Austin Lee Nichols and Jon K Maner. The good-subject effect: Investigating participant demand characteristics. *The Journal of general psychology*, 135(2): 151–166, 2008.

Hiroko Nishida. Cultural schema theory. *Theorizing about intercultural communication*, 401418, 2005.

Robert M Nosofsky, Craig A Sanders, and Mark A McDaniel. Tests of an exemplar-memory model of classification learning in a high-dimensional natural-science category domain. *Journal of Experimental Psychology: General*, 147(3):328, 2018.

Martin A Nowak and Robert M May. Evolutionary games and spatial chaos. *Nature*, 359(6398):826–829, 1992.

Karine Nyborg, John M Anderies, Astrid Dannenberg, Therese Lindahl, Caroline Schill, Maja Schlüter, W Neil Adger, Kenneth J Arrow, Scott Barrett, Stephen Carpenter, and Others. Social norms as solutions. *Science*, 354(6308):42–43, 2016.

Linda M O'Keeffe, Gemma Taylor, Rachel R Huxley, Paul Mitchell, Mark Woodward, and Sanne AE Peters. Smoking as a risk factor for lung cancer in women and men: a systematic review and meta-analysis. *BMJ open*, 8 (10):e021611, 2018.

Andreas Ortmann. Deception. In *Handbook of research methods and applications in experimental economics*. Edward Elgar Publishing, 2019.

James Otteson. Adam smith's first market: the development of language. *History of Philosophy Quarterly*, 19(1):65–86, 2002.

Hye-Jin Paek and Thomas Hove. How the media effects schema and the persuasion ethics schema affect audience responses to antismoking campaign messages. *Health communication*, 33(5):526–536, 2018.

Maria Pia Paganelli. The adam smith problem in reverse: Self-interest in the wealth of nations and the theory of moral sentiments. In *Theory and practice of economic policy. Tradition and change. Selected Papers from the 9th Aispe Conference: Tradition and change. Selected Papers from the 9th Aispe Conference*, page 73. FrancoAngeli, 2009.

Andrea L Patalano, Edward E Smith, John Jonides, and Robert A Koeppe. Pet evidence for multiple strategies of categorization. *Cognitive, Affective, & Behavioral Neuroscience*, 1(4):360–370, 2001.

Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior research methods*, 46(4):1023–1031, 2014.

Julia Penelope. *Speaking freely: Unlearning the lies of the father's tongues*. Perfamon, 1990.

Christina Petsoulas. *Hayek's liberalism and its origins: His idea of spontaneous order and the Scottish Enlightenment*. Routledge, 2013.

Alexander Peysakhovich and David G Rand. Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, 62(3):631–647, 2015.

Alexander Peysakhovich and David G Rand. Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, 62(3):631–647, 2016.

Alban William Housego Phillips. *Employment, inflation and growth*. Bell, 1962.

Pablo Pirnay-Dummer, Dirk Ifenthaler, and Norbert M. Seel. Semantic networks. In Norbert M. Seel, editor, *Encyclopedia of the Sciences of Learning*, pages 3025–3029. Springer US, Boston, MA, 2012. ISBN 978-1-4419-1428-6. doi: 10.1007/978-1-4419-1428-6_1933. URL https://doi.org/10.1007/978-1-4419-1428-6_1933.

Tadeusz Platkowski. Enhanced cooperation in prisoner's dilemma with aspiration. *Applied mathematics letters*, 22(8):1161–1165, 2009.

Karl Popper. *The poverty of historicism*. Routledge, 2013, 1957.

Karl Popper. *Conjectures and refutations: The growth of scientific knowledge*. routledge, edition: 2014, 1963.

Thomas Pownall. *A letter from Governor Pownall to Adam Smith*. Sentry Press, 1776.

John Protzko, Brett Ouimette, and Jonathan Schooler. Believing there is no free will corrupts intuitive cooperation. *Cognition*, 151:6–9, 2016.

Matthew Rabin. Incorporating fairness into game theory and economics. *The American economic review*, pages 1281–1302, 1993.

David G Rand. The promise of mechanical turk: How online labor markets can help theorists run behavioral experiments. *Journal of theoretical biology*, 299:172–179, 2012.

Kate Raworth. *Doughnut economics: seven ways to think like a 21st-century economist*. Chelsea Green Publishing, 2017.

David Ricardo. *An essay on the influence of a low price of corn on the profits of stock, with remarks on mr. Malthus' two last publications*. John Murray, London, 1815.

David Ricardo. On the principles of political economy and taxation. *publicado en*, 1817.

George D Romanos. Reflexive predictions. *Philosophy of Science*, 40(1):97–109, 1973.

Lee Ross, David Greene, and Pamela House. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, 13(3):279–301, 1977.

Myron Rothbart. Memory processes and social beliefs. *Cognitive processes in stereotyping and intergroup behavior*, pages 145–181, 1981.

Jean-Jacques Rousseau. *Rousseau: The Social Contract and other later political writings*. Cambridge University Press, 2018, 1762.

Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica: Journal of the Econometric Society*, pages 97–109, 1982.

Ariel Rubinstein. A sceptic's comment on the study of economics. *The Economic Journal*, 116(510):C1–C9, 2006.

Warren J Samuels. Adam smith's invisible hand. *The Street Porter and the Philosopher: Conversations on Analytical Egalitarianism*, pages 179–201, 2008.

Warren J Samuels. *Erasing the invisible hand: Essays on an elusive and misused concept in economics*. Cambridge University Press, 2011.

Paul Anthony Samuelson. *Economics: an introductory analysis*, volume 715. McGraw-Hill New York, 1948a.

Paul Anthony Samuelson. Foundations of economic analysis. *Science and society 13(1):93-95*, 1948b.

Theo GM Sandfort. Sexual orientation and gender: Stereotypes and beyond. *Archives of Sexual Behavior*, 34(6):595–611, 2005.

William H Sandholm. Evolutionary game theory. *Encyclopedia of Complexity and Systems Science*, pages 3176–3205, 2009.

Ana C Santos. Behavioural and experimental economics: are they really transforming economics? *Cambridge Journal of Economics*, 35(4):705–728, 2011.

Ana C Santos and João Rodrigues. Economics as social engineering? questioning the performativity thesis. *Cambridge Journal of Economics*, 33(5): 985–1000, 2009.

Francisco C Santos and Jorge M Pacheco. Scale-free networks provide a unifying framework for the emergence of cooperation. *Physical Review Letters*, 95(9):098104, 2005.

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

Barry Schwartz, Richard Schuldenfrei, and Hugh Lacey. Operant psychology as factory psychology. *Behaviorism*, pages 229–254, 1978.

Reinhard Selten and Axel Ockenfels. An experimental solidarity game. *Journal of economic behavior & organization*, 34(4):517–539, 1998.

Amartya Sen. Economics, business principles and moral sentiments. *Business Ethics Quarterly*, 7(3):5–15, 1997.

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

Olivier Serrat. Social network analysis. In *Knowledge solutions*, pages 39–43. Springer, 2017.

Leigh S Shaffer. Toward pepitone's vision of a normative social psychology: What is a social norm? *The journal of mind and behavior*, pages 275–293, 1983.

Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118, 1955.

Herbert A Simon. Theories of bounded rationality. *Decision and organization*, 1(1):161–176, 1972.

Brian Skyrms. The stag hunt. In *Proceedings and Addresses of the American Philosophical Association*, volume 75, pages 31–41. JSTOR, 2001.

Dirk Smeesters, S Christian Wheeler, and Aaron C Kay. Indirect prime-to-behavior effects: The role of perceptions of the self, others, and situations in connecting primed constructs to social behavior. In *Advances in experimental social psychology*, volume 42, pages 259–317. Elsevier, 2010.

Adam Smith. *The theory of moral sentiments*. Penguin Group, 2009, 1759.

Adam Smith. *Wealth of Nations: A selected edition*. Oxford UP, ed. Sutherland, K, 2008, 1776.

David Smith. *Free lunch: Easily digestible economics*. Profile Books, 2010.

Chun Siong Soon, Marcel Brass, Hans-Jochen Heinze, and John-Dylan Haynes. Unconscious determinants of free decisions in the human brain. *Nature neuroscience*, 11(5):543, 2008.

Dugald Stewart et al. Account of the life and writings of adam smith. *Histoy of Economic Thought Chapters*, 10:1–98, 1793.

Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. Measuring large-scale social networks with high resolution. *PloS one*, 9(4):e95978, jan 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0095978. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0095978.

Susan Strange. *States and markets*. Bloomsbury Publishing, 2015.

Cass R Sunstein. Nudging: a very short guide. *Journal of Consumer Policy*, 37 (4):583–588, 2014.

Lewis Madison Terman. *Sex and personality: Studies in masculinity and femininity*. Yale University Press, 1936.

WI Thomas and DS Thomas. *The child in America.* A. A. Knopf, 1928.

Marco Tomassini, Enea Pestelacci, and Leslie Luthi. Social dilemmas and cooperation in complex networks. *International Journal of Modern Physics C*, 18(07):1173–1185, 2007.

Colin Turnbull. *Mountain people*. Simon and Schuster, 1987.

Amos Tversky and Daniel Kahneman. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

Amos Tversky and Daniel Kahneman. Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics*, 106(4): 1039–1061, 1991.

Christian Unkelbach, Alex Koch, Rita R Silva, and Teresa Garcia-Marques. Truth by repetition: Explanations and implications. *Current Directions in Psychological Science*, 28(3):247–253, 2019.

E Van Avermaet. *Equity: A Theoretical and Empirical Analysis*. Unpublished doctorial dissertation, University of California, Santa Barbara, 1974.

Rob Van Horn and Philip Mirowski. The rise of the chicago school of economics and the birth of neoliberalism. *The Road from Mont Pelerin: The Making of the Neoliberal Thought Collective. Harvard University Press: Boston, MA*, pages 149–163, 2009.

Robert Van Horn. Hayek and the chicago school. In *Hayek: A Collaborative Biography*, pages 91–111. Springer, 2015.

Hal R Varian. *Intermediate microeconomics with calculus: a modern approach*. WW Norton & Company, 2014.

Jacob Viner. Adam smith and laissez faire. *Journal of political economy*, 35(2): 198–232, 1927.

Kathleen D Vohs and Jonathan W Schooler. The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological science*, 19(1):49–54, 2008.

Francois Voltaire. *Letters on England*. Penguin UK, 1980, 1733.

John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior (commemorative edition)*. Princeton university press, 2007, 1944a.

John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1944b.

Willard Waller and Reuben Hill. *The family: A dynamic interpretation.* Dryden Press, 1951, 1938.

Léon Walras. Elements of pure economics. *Translated from the French by William Jaffé, 1954*, 1874.

Michael Walzer. Passion and politics. *Philosophy & social criticism*, 28(6): 617–633, 2002.

Long Wang, Deepak Malhotra, and J Keith Murnighan. Economics education and greed. *Academy of Management Learning & Education*, 10(4):643–660, 2011.

Gregory Wheeler. Bounded rationality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2020 edition, 2020.

Jessica Whyte. The invisible hand of friedrich hayek: Submission and spontaneous order. *Political Theory*, 47(2):156–184, 2019.

Robert H Wicks. Schema theory and measurement in mass communication research: Theoretical and methodological issues in news information processing. *Annals of the International Communication Association*, 15(1): 115–145, 1992.

Jonathan B Wight. Antecedents to the crisis: Mandeville, smith, and keynes. *International Journal of Social Economics*, 2018.

Nick Wilkinson and Matthias Klaes. *An introduction to behavioral economics*. Macmillan International Higher Education, 2017.

Marc Willinger and Anthony Ziegelmeyer. Framing and cooperation in public good games: an experiment with an interior solution. *Economics letters*, 65(3):323–328, 1999.

James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

James Woodward. Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, 25(3):287–318, 2010.

Zhi-Xi Wu, Jian-Yue Guan, Xin-Jian Xu, and Ying-Hai Wang. Evolutionary prisoner's dilemma game on Barabási–Albert scale-free networks. *Physica A: Statistical Mechanics and its Applications*, 379(2):672–680, jun 2007. ISSN 0378-4371. doi: 10.1016/J.PHYSA.2007.02.085. URL https://www.sciencedirect.com/science/article/pii/S0378437107001677.

Robert Wuthnow. Acts of compassion: Caring for ourselves and helping others. *Princeton, NJ*, 1991.

Alison Wylie and Lynn Hankinson Nelson. Coming to terms with the value(s) of science: Insights from feminist science scholarship. In Harold Kincaid, John Dupre, and Alison Wylie, editors, *Value-Free Science? Ideals and Illusions*, pages 58–86. Oxford University Press, Usa, 2007.

Erte Xiao and Cristina Bicchieri. When equality trumps reciprocity. *Journal of Economic Psychology*, 31(3):456–470, 2010.

Toshio Yamagishi, Karen S Cook, and Motoki Watabe. Uncertainty, trust, and commitment formation in the united states and japan. *American Journal of Sociology*, 104(1):AJSv104p165–194, 1998.

Anthony M Yezer, Robert S Goldfarb, and Paul J Poppen. Does studying economics discourage cooperation? watch what we do, not what we say or how we play. *Journal of Economic Perspectives*, 10(1):177–186, 1996.

Jeffrey T Young. *Elgar Companion to Adam Smith*. Edward Elgar Cheltenham, 2009.

Michel S Zouboulakis. From mill to weber: the meaning of the concept of economic rationality. *European Journal of the History of Economic Thought*, 8 (1):30–41, 2001.

Appendix

Self-interested behaviour as a social norm

Kamilla Haworth Buchter[1◕], Bjarke Mønsted[2◕], Sune Lehmann[2*]

1 London School of Economics, department of philosophy, logic, and scientific method, London, United Kingdom

2 Technical University of Denmark, department of applied mathematics and computer science, 2800 Kgs. Lyngby, Denmark

◕These authors contributed equally to this work.

* sljo@dtu.dk

Abstract

Language can exert a strong influence on human behaviour. In experimental studies, it is for example well-known that the framing of an experiment[1] or priming at the beginning of an experiment[2] can alter participants' behaviour. However, few studies have been conducted to determine why framing or priming specific words can alter people's behaviour[3, 4]. Here, we show that the behaviour of participants in a game-theoretical experiment is driven mainly by social norms[5], and that participants' adherence to different social norms is influenced by the exposure to economic terminology. To explore how these terminology-driven changes impact behavior at the system level, we use established frameworks for modeling collective cooperative behaviour[6, 7]. We find that economic terminology induces a behavioural difference which is larger than that caused by financial incentives in the magnitude usually employed in experiments and simulation. These findings place an increased responsibility on scientists and science communicators, as scientific terminology is increasingly communicated to the general population[8, 9, 10].

Introduction

We start from the observation that economists tend to exhibit substantially different behaviour in experiments pertaining to game theory and economics[3]. In particular, those who study, or have studied, economics tend to act more in accordance with the predictions of microeconomic theory, acting to maximize their own profits[4]. This raises the question why economists behave differently. Some have proposed a self-selection mechanism, according to which people more inclined to maximize their own profits are also more likely to choose to study economics[11]. Others have proposed a learning mechanism where exposure to economic theory is the cause of the behavioural differences[12]. In the past ten years, studies have repeatedly confirmed the latter mechanism where

changes in behaviour are caused by exposure to microeconomic theory[13]. However, few studies have engaged with the question of why a learning effect occurs[3, 14, 15, 4].

Here, we present evidence that engaging with microeconomic terminology inhibits cooperative behaviour in a competitive game setting, and that engaging with an alternative terminology which emphasizes collective, rather than individual, payouts increases cooperative behaviour. Applying Bicchieri's definition of social norms[5], we next show that these terminology-driven changes are caused by social norms and not by alternative effects such as the terminologies enabling participants to understand the experiment or biasing participants to prefer certain behaviours. We finally use simulation methods from evolutionary game theory to assess whether observed individual behavioural differences arising from terminology exposures are sufficient to take a collective system across a tipping point to states of complete cooperation or defection in a simulated population. The findings have impact beyond economics since they suggest that scientific terminology can guide people's behaviour by prompting specific social norms and that behaviour guided by these norms can completely determine the outcome of a collective system.

For the initial experiments, we built an online platform where participants could play 10 rounds of prisoner's dilemma game (PDG). In a PDG, participants are faced with a choice to either *defect* or *cooperate* with another player. If both players cooperate, their combined payout will be maximized. However, for each individual player, choosing defection will maximize their own payout, regardless of the choice of the other player. Finally, if both players' defect, they will each receive a smaller payout compared to the payout they get if both players cooperate. In order to test the behavioural effects of scientific terminology, participants were randomly assigned to one of three categories - *individualist* (I), *collectivist* (C), or *neutral* (N). Before proceeding to the game, participants in all three groups had the structure of the PDG explained to them, and were required to correctly answer a series of control questions to ensure they understood how their payout depended on the actions of both players. In addition, participants in (C) and (I) were introduced to two distinct concepts from microeconomics, and asked to apply them when answering the control questions, and throughout the PDG. Participants in (I) were shown introduced to the microeconomic concept of *rationality* and explained that in game theory it is called *rational* to maximize ones own reward by defecting[16]. In similar fashion, participants in (C) were introduced to the concept of *social optimality*[17] and explained that in game theory it is called *optimal* to maximize collective wealth by cooperating. Players in (N) were not introduced to any concepts and their five control questions emphasized collective and individual gain to the same extent.

Having tested for the behavioural effects of terminologies, we subsequently ran a follow-up experiment to understand the role of social norms in the decision process. Here, participants were asked two questions before playing each PDG. The first question asked which move they expected the other player to make. The second questions asked which move they thought, the other player expected them to make. Our goal was to determine whether there is a connection between

participants' expectations and their behaviour and whether the behavioural effects caused by the terminologies are driven by these expectations such that they are caused by adherence to a social norm of cooperation or to a social norm of defection[5].

If terminology can substantially change which social norms individuals follow, the question then becomes: What is the effect of terminology at the collective level? We explore this question through the lens of evolution of cooperative behavior[18, 19, 20]. Specifically we simulate agents embedded in a network and use decision heuristics informed by empirical data on terminology-driven changes in social norms to decide agents' actions in a PDG based on their surroundings, following [21, 22, 23, 24, 25]. We run these simulation experiments on artificial networks commonly used in the literature, as well as networks constructed from real world data, using data from the Copenhagen Networks Study[26].

Terminology influences behaviour

In order to assess the interplay between terminology and educational background, we recruited participants in the (Behavioral Research Lab (BRL)) at the London School of Economics (LSE) ($n = 462$) and on Amazon's Mechanical Turk (MTurk) ($n = 344$). The participants played a PDG with 10 rounds on an online platform, through which they were informed they would play against a new participant in each round. In reality, they were playing a computer choosing randomly between defection and cooperation. The experimental setup was approved by the LSE Research Ethics Committee.

Among the BRL participants, 77 were associated with a degree involving at least two years with economics courses. The degrees were core degrees from the three departments Economics, Finance, and Accounting, and we refer to participants associated with these degrees as *economists* for short. Consistent with results from previous PDG experiments[12, 13], we find that economists cooperate less than the remaining BRL participants (linear regression on number of cooperate moves, $p = .0007$, $n = 462$, $t = -3.22$, one-tailed). Comparing data sources, BRL participants cooperated significantly less ($p = .048$, $n = 810$, $t = -1.66$, one-tailed). However, this difference disappears when excluding economists from the analysis ($p = .25$, $n = 733$, $t = -0.66$, one-tailed).

Focusing on the remaining 733 participants from BRL and MTurk, we compare the number of times participants in the collectivist, neutral, and individualist categories defected in the 10 rounds (fig. 1a). A Kruskal-Wallis test revealed a significant interaction between terminology and cooperation ($p = 6.8 \cdot 10^{-9}$, $n = 733$, $H = 38$). A Conover-Iman post-hoc analysis[27] showed that, compared with the neutral category, participants exposed to the collectivist terminology cooperated more ($p = .03$, $n = 487$, $t = 1.89$, , one-sided), while the individualist category defected more ($p = 8.2 \cdot 10^{-6}$, $n = 501$, $t = 4.34$, one-sided). Throughout the 10 rounds, participants in all categories became more likely to defect (fig. 1b).

Considering the BRL participants, the difference between economists and non-economists is consistent across all categories (fig. 1c-e). The magnitude of this behavioural difference between

economists and non-economists is comparable to the effect of economic terminology. In particular, economists exposed to a neutral terminology defected as much as regular participants exposed to the individualist terminology (fig. 1d-e).

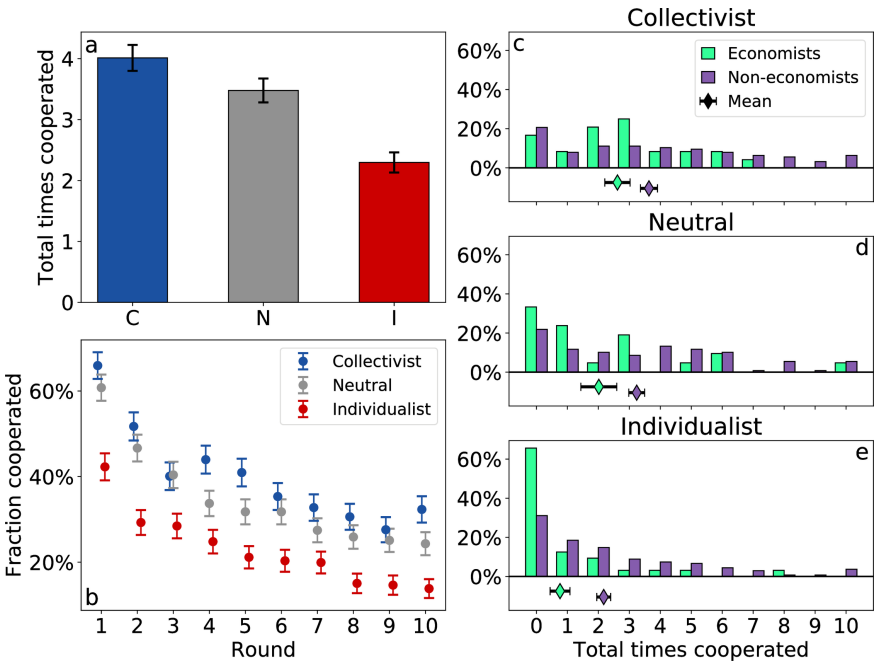


Fig 1. Interaction between terminology and behaviour. **a** Participants who had the game described in collectivist (C) or individualist (I) terms cooperated and defected more, respectively. **b** throughout the 10 rounds, participants in all categories became increasingly more likely to defect. Some difference persisted over time, especially individualists defecting more than others. **c-e** Distributions over the number of cooperation moves among participants exposed to the three terminologies, grouped by whether participants had a background in economics. Bars represent the observed frequencies, and the markers represent the mean number of times cooperated. Participants affiliated with an economics-related discipline consistently cooperated fewer times than other participants, but terminology exposure had a similar influence on both groups. The mean for each distribution is indicated with black markers. Error bars in all subfigures represent the standard error of the means.

Behavioural changes are mediated by social norms

One potential explanation of why terminology influences behaviour, is that language impacts the salience of different social norms[14, 4]. We conducted a second experiment on MTurk ($n = 200$) to understand whether social norms provide an explanation of the observed behavioural effects.

We adopt Bicchieri's definition of social norms[5], which views a person's behaviour is expressive of a social norm if 1) they are aware that a behavioural rule exists and applies to their situation, and 2) the person's conforming to the rule is contingent on their first and second order beliefs regarding general compliance, i.e. they must generally expect others to comply with the behavioural rule, and believe others to expect them to comply, too. Note that conflicting norms can exist under this definition[5]. The terminologies in the three categories were designed to provide participants

with cues that a behavioural rule of cooperation (defection) exists for the collectivist (individualist) terminologies, see methods for details.

To assess beliefs regarding compliance, participants were asked before each round whether they expected the other player to defect or cooperate, and which choice they believed the other player expected of them in turn. In the following, we will use the phrase 'expecting cooperation (defection)' as shorthand for participants who expect cooperation (defection) of their opponent and believes the same is expected of them. Note that the two are not exhaustive, as the first and second order beliefs need not align. If the mechanism through which terminology influences behaviour is social norms of cooperation and defection, we should expect any behavioural effect to be strongly contingent on beliefs about compliance[28].

The results from the second experiment indeed show a strong correlation between expectations and behaviour, as shown in fig. 2a. For example, in the first round, 88.2% of participants expecting cooperation chose to cooperate, whereas 11.2% of those expecting defection did so. Within the subgroups of participants expecting cooperation and defection, fig. 2a shows only a slight difference between participants exposed to the three terminologies. However, terminology exposure significantly impacts the probability for participants to hold such expectations, as depicted in fig. 2b. Specifically, participants in (C) were significantly more likely to expect cooperation ($p = .02$, $n = 133$, $z = 2.06$, one-tailed proportional z-test), and participants in (I) significantly less so ($p = .04$, $n = 145$, $z = -1.78$, one-tailed proportional z-test), when compared to the neutral (N) group. The number of times participants expected cooperation over the ten rounds also varied significantly with terminology exposure ($p = .003$, $n = 200$, $H = 12$, Kruskal-Wallis test), and similarly for defection expectations ($p = .044$, $n = 200$, $H = 6.3$). Excluding participants who responded in a post-experiment survey that they either felt compelled to cooperate or defect more, or suspected that the aim of the experiment was to influence cooperative behaviour, did not significantly influence the results in this or the first experiment (details in methods).

The results visualized in fig. 2 support the hypothesis that scientific language, specifically microeconomic terminology, can influence people's behaviour by encouraging them to follow different social norms. We draw this conclusion because the results cannot be adequately explained by the alternative hypotheses. If the terminologies affected behaviour by biasing participants to prefer one action over the other, participants' choices would not depend on their first and second order expectations regarding other participants[2]. Another proposed mechanism for such behavioural effects is that language may shift the underlying utility functions for participants, such that acting e.g. in an altruistic fashion results in a higher utility in spite of the lower payout[15]. This explanation is also not compatible with the results, as participants were required to answer control questions to ensure their understanding that defection (cooperation) would maximize individual (collective) payout *regardless* of the choice of the other player. Therefore, if participants were simply maximizing an underlying utility with larger values for individualist or altruistic behaviour, beliefs

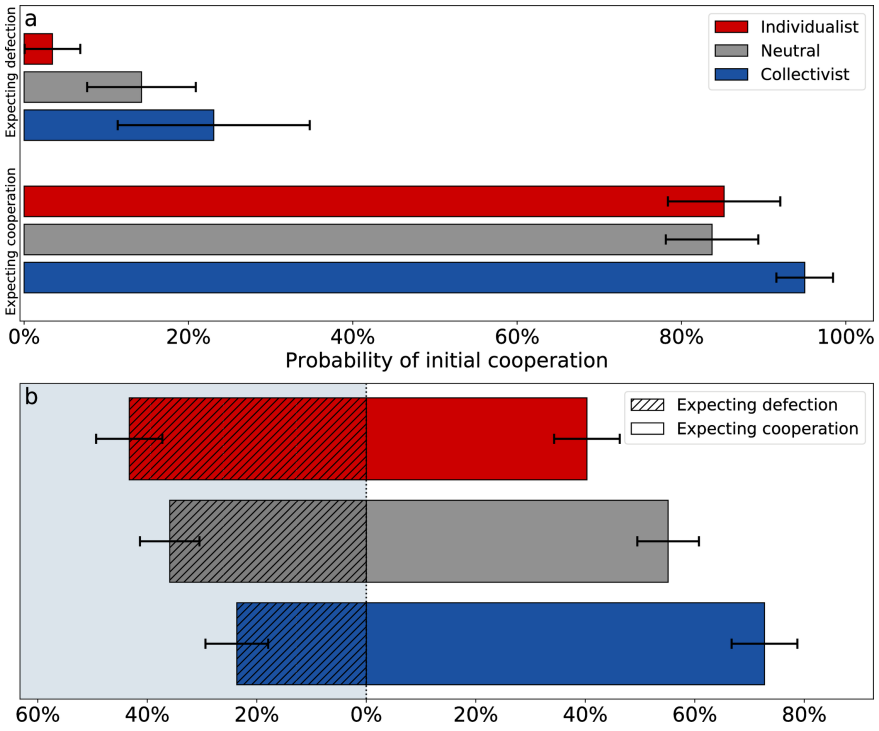


Fig 2. The interplay between expectations and terminology exposure. **a** The percentage of participants choosing to cooperate in each round, grouped by expectations and terminology exposure. When participants expect cooperation, meaning expecting the other player to cooperate and believing the other player expects them to cooperate, they are much more likely to cooperate. Similarly, participants who expect defection are much more likely to defect. **b** The percentage of participants exposed to each of the three terminologies (collectivist, neutral, and individualist) who expect cooperation and defection, respectively. Participants exposed to the individualist terminology become more likely to expect others to defect, and to believe that others in turn expect defection from them, and less likely to hold the similar beliefs for cooperation. The opposite effect is seen for exposure to the collectivist terms. The error bars represent error of the means.

regarding general compliance should not affect the choices of participants. For the same reason, it cannot be the case that terminology exposure simply helps participants better understand how to maximize a preexisting individualist or altruist utility function.

The behavioural effects of terminology can drive simulated collective systems across tipping points

The experimental results show that terminology can influence individual behaviour in a PDG and that this influence can be explained in a social norm framework by a change in beliefs about other people's intentions and expectations. Having established the impact on social norms at the individual level, we now explore how such terminology-driven shift in norms may impact behavior on the systemic level. It has been shown that individual behavioural differences caused by social norms may drive a collective system across a tipping point[29] and cause dramatic collective effects. To investigate whether the observed individual behavioural effect from exposure to microeconomic terminology are sufficient to drive a collective system across a tipping point in our case, we use established methods from agent-based evolutionary game theory[21, 30] to simulate participants exposed to the various terminologies interacting with each other in a network.

We run simulations on a real-world interaction network, obtained from smartphone data from over 700 students in the Copenhagen Networks Study[26]. We construct interaction networks in which nodes represents students, and links represent interactions along several channels including text messages, physical proximity (measured by Bluetooth), and Facebook friendships.[31]. We report results from the Facebook friendship network here, and refer to the SI for similar results using the remaining networks, as well as commonly used artificial networks.

Agents in the simulation employ stochastic update heuristics[32], specifically a biased logit model[22],

$$p_c = \frac{1}{1 + \mathrm{e}^{-\beta \mathbf{w} \cdot \mathbf{x}}}, \tag{1}$$

which we fitted to the experimental data. In eq. (1), $\mathbf{w}$ is a weight vector including bias terms, which were adjusted to the maximum likelihood fit to the experimental data. $\mathbf{x}$ is a state vector which denotes quantities such as the past moves of the agent and their last opponent.

As participants in the three categories (C, N, and I) exhibited quite different behaviours, we repeated this procedure independently for each group, and thus obtained three distinct agent-level decision heuristics.

In order to assess any emergent phenomena the induced individual behavioural differences might give rise to, we ran a series of simulated repeated PDGs. In order to probe specifically the induced behavioural differences, we subtracted the bias terms from the neutral model from the remaining two, so the resulting models represented the difference in behaviour relative to the neutral category. Details of this, along with parameter values, fit quality measures, details on alternate models, and visualizations of the resulting heuristics, are available in the SI. We then defined a parameter ρ_I denoting the fraction of agents in a simulation which used the decision

heuristic based on experimental data in the individualist category. The remaining $1 - \rho_I$ fraction of agents would then act according to the model fitted to data from the collectivist category. Results from simulations on the real-world Facebook friendship network are illustrated in fig. 3.

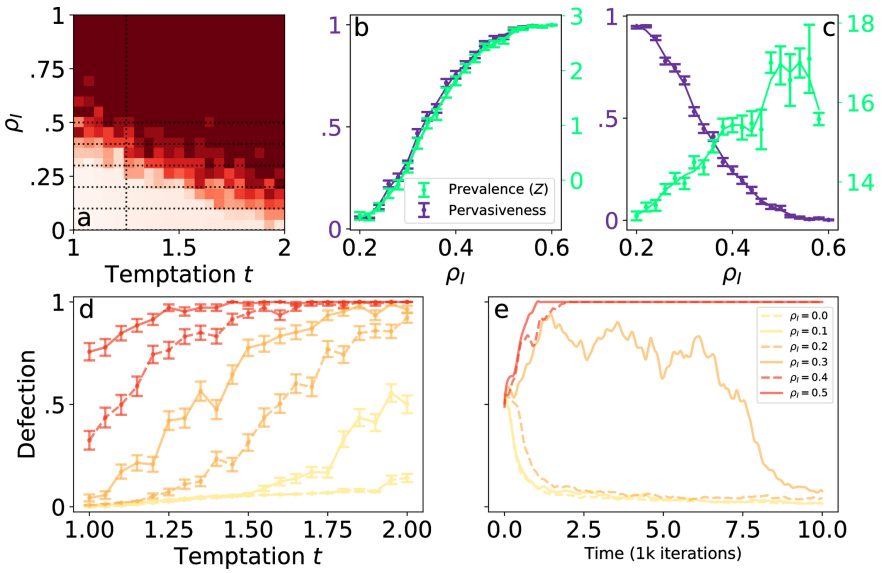


Fig 3. Simulation results for a range of values of the 'temptation to defect' t, and the proportion ρ_I of simulated individuals which act similarly to people exposed to individualist, rather than collectivist, terminology. **a** The tendency of cooperation to disappear entirely in simulation, for a range of parameter values. Predictably, this tendency increases with the temptation to defect. However, varying the ratio of collectivists vs. individualists in the simulation has much more pronounced effects, and takes the system across a tipping point, from complete cooperation to complete defection. **b, c** Commonality measures for defection and cooperation, respectively, as a function of the temptation parameter. Global commonality (pervasiveness) increases and decreases, respectively, with temptation. Local commonality (prevalence), however, increases with ρ_I for both strategies, as the network becomes more polarized and cooperators are forced to be more tightly clustered together in order to survive in spite of the growing number of defectors. **d** Defection rates for selected values of ρ_I (indicated with dashed lines in **a**), averaged over the final 5k iterations of simulations, as a function of temptation. **e** Progression of individual simulations for specific values of ρ_I and t (indicated by the intersections of lines in **a**). The system generally goes to complete cooperation or defection, but is volatile for values around $\rho_I = .3$. The lines are smoothened using a Savitzky–Golay filter with window length 400 and polynomial order 5. All error bars represent the error of the means.

In these simulations, results were more sensitive to changes in ρ_I compared to t. Scanning the temptation parameter t across the $[1, 2]$ range gradually increases overall defection rates, while increasing ρ_I takes the entire network from complete, or almost complete, cooperation, across a tipping point into a regime of complete defection, regardless of t. The proportion of simulations in which cooperators died out entirely is shown for a range of parameter values in fig. 3a.

For selected values of ρ_I, the outcomes of simulations for varying temptation values are shown in fig. 3b. For the same values of ρ_I, and a single value of t, the progression of a single simulation over $10,000$ rounds is displayed in fig. 3c, showing that for high and low ρ_I, the system quickly settles to states of complete defection and cooperation, respectively. For values in between, the system is volatile and can go either way.

While increasing ρ_I leads to decreased cooperation overall, the cooperators that remain tend to

be more tightly clustered together. We can visualize this by dividing the notion of commonality into *pervasiveness* (global) and *prevalence* (local). We define the pervasiveness of a behaviour (cooperation or defection) simply as the fraction of nodes in a network that partake in that behaviour at the end of a simulation. Prevalence is defined as the z-score of the observed number of e.g. cooperate-cooperate neighbours, compared to a random permutation null model. Figure 3d shows the commonality measures for defection as ρ_I is increased, and fig. 3e shows the same for cooperation, showing that cooperators become much more tightly nit together (increasing prevalence) as their numbers dwindle (decreasing pervasiveness).

Discussion

Our findings indicate that cooperative behaviour may be significantly influenced by exposure to scientific terminology. Compared with a neutral group, we found that using different, but equivalent, scientific terms to describe a competitive experiment, we could both amplify and dampen cooperative behaviour. The terminology which reduced cooperative behaviour is standard microeconomic terminology, and the reduction in cooperative behaviour was comparable to the difference observed between participants enrolled in educations with and without a heavy background in economics. We saw strong evidence that people's choice to cooperate or defect were in part governed by social norms, as participants were much more likely to elect a move which they expected their opponent to play, and which they believed their opponent in turn expected of them. The terminologies were able to alter behaviour in our participants, by manipulating these expectations. We then used simulations to show that the behavioural differences introduced by different terminologies were sufficient to drive a system across a tipping point between states of complete cooperation and defection.

As science is becoming more broadly and popularly disseminated in the population, this places a greater responsibility on science communicators to be understand that scientific language may contain value laden terms which interact with social norms to produce emergent behavioural changes. In addition, the findings emphasize the need for the experimental behavioural scientist to be aware of how observed behaviours might be influenced by familiarity with relevant scientific terminologies, both from the explanation of the experiment received by participants, and due to popular-scientific communication. On a grander canvas, our results highlight how, for example, the language of leaders or media – choosing to focusing on selfish rationality or social cooperation – may drive real changes in people's behavior.

In future experiments, it would be interesting to look for similar interactions between norms and terminology within other scientific disciplines, as well as to expose groups of participants to the same terminological stimulus to investigate directly the degree to which terms give rise to emergent effects. Finally, further experimentation might attempt to directly influence empirical and normative expectations, by providing participants with direct evidence that a certain behaviour,

e.g. cooperation or defection, is more probable from other participants, or that other participants are more likely to expect said behaviour.

Materials and methods

Experimental setup and data collection

The initial experiment on MTurk was carried out in November 2018. We used a predefined setting on the MTurk platform to allow only 'master' workers, with a consistent history of delivering high quality work, to assess the experiment. Workers were paid a base pay of 2 USD for participating, plus 10% of the points gained throughout the experiment as USD, resulting in a median earning of 3.6 USD. A total of 344 workers participated.

The experiment was repeated in December 2018, in the BRL at the LSE, with 466 participants completing the experiment in batches of up to 20 people over the course of a week. 4 participants were excluded for various reasons - 1 did not have a valid birth year in the laboratory's database, 1 did not identify as either male or female, and 2 had opted not to fill out a short post-experiment questionnaire in which participants could indicate whether they had played a similar game previously.

Here, participants where required to be physically present in the laboratory, playing the PDG on computers located in individual booths. Participants received a base pay of 5 GBP plus another 10% of the points obtained in the experiment in GBP, resulting in a median earning of 6.7 GBP. Of the 466 participants, 77 associated with a degree in either economics, finance, or accounting, each of which requires at least 2 years with economics courses.

Finally, 200 participants were again recruited on MTurk in March 2019 to conduct the follow-up experiment in which inquired about participants' expectations regarding the choices of other players, as well as their beliefs about the expectations of others. The workers were compensated in a similar fashion as in the earlier MTurk experiment, with a median pay of 2.8 USD.

Players in the three categories were provided with different *situational cues* in the experimental description by referring to other participants as either *your opponent* (I), *the other participant* (N), or *your co-player* (C). This was intended to highlight competitive or communal aspects of the game (fig. 4a).

Players were not allowed to proceed to the PDG before they had answered all five control questions correctly. In this way we ensured that players understood the game and the terminologies we had introduced. In addition, players would be asked to apply the learned terminology after each of the rounds, to ensure their continued engagement (fig. 4c). More details and screenshots of the platform are provided in the SI.

The additional text and focus in the control questions in (I) and (C) provided players with normative cues, by applying value-laden labels to the defect and cooperation strategies (fig. 4a).

Assessing experimenter bias

In order to mitigate effects of experimenter bias, participants were asked - after completing the experiment - whether they suspected that a particular hypothesis was being tested. Participants

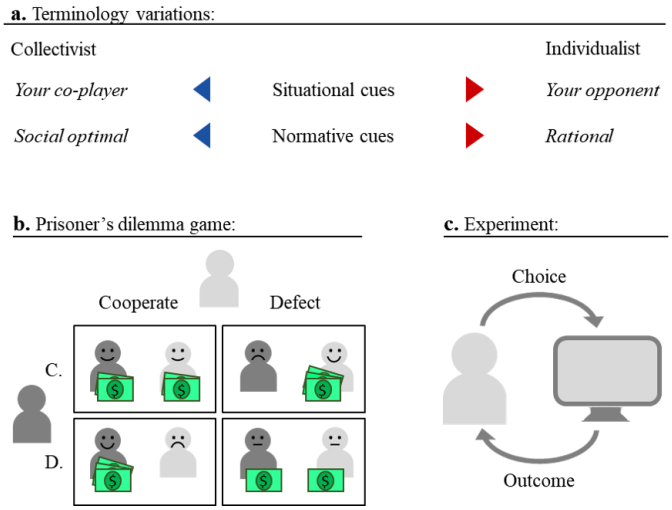


Fig 4. Overview of the experimental setup. **a** Participants in the collectivist (C) and individualist (I) categories receive situational cues emphasizing competitive or communal aspects of the game. Participants in the neutral (N) category received no normative cues and other participants were referred to as "the other participant" to avoid situational cues. **b** The structure of a prisoner's dilemma game. If players cooperate, their combined payout is maximized. Each individual player will receive a larger playout from defecting rather than cooperating, independently of the other player's action. **c** Participants answer a series of questions to ensure they understand the game and the terminology being introduced to them. In each round, they choose to cooperate or defect, and the computer presents them with the results of the present round, and asks them a follow-up question involving the introduced terminology to ensure their continued understanding and engagement.

who responded positively were allowed to describe in a text field what they believed was being tested in the experiment. We identified responses which indicated that participants either believed that some effect of language on behaviour was being tested, or that a particular action was in some way the right one. A summary of the responses is shown in table 1.

	BRL			MTurk 1			MTurk 2		
	C	N	I	C	N	I	C	N	I
"Language"	6	0	7	1	1	2	2	0	1
"Defect"	13	12	24	2	2	3	0	0	0
"Cooperate"	8	4	3	0	1	2	1	0	0

Table 1. Overview of the number of people who reported in a post-experiment questionnaire to suspect that various effects were being tested. "Language" means the participant stated that they believed the aim of the experiment was to test some form of effect of language on participant behaviour. "Defect" and "cooperate" refers to participants believing that defection or cooperation was the right strategy to choose.

Redoing the statistical analyses while excluding participants from table 1, the main findings remained significant. The number of times participants cooperated was still different in the three terminology groups ($p = 1.4 * 10{-}8$, $n = 660$, $H = 36$, Kruskal-Wallis test). The fraction of participants expecting cooperation (believing their opponent would cooperate, and believing the same was expected of them) in the first round was also higher for the collectivist terminology group ($p = .017$, $n = 192$, $z = 2.11$, logistic regression, one-tailed z-test), but not so for defection ($p = .15$, $n = 192$, $z = 1.03$). However the total number of times participants expected cooperation/defection

across the ten rounds remained significantly higher in the collectivist/individualist exposure groups.

Models and parameter estimation

We model the choices of an individual agent as a stochastic function of variables representing information available to the agent. We considered the family of logit dynamics[22] models, which has previously been used in the context of evolutionary game theory[32]. We write a general logit dynamics model on the form

$$\begin{aligned} p_i &= \frac{\mathrm{e}^{\beta \mathbf{w}_i \cdot \mathbf{x}}}{Z}, \\ Z &= \sum_j \mathrm{e}^{\beta \mathbf{w}_j \cdot \mathbf{x}}. \end{aligned} \tag{2}$$

Here, $\mathbf{x}$ is an input vector representing the information based on which an agent makes their decision, and $\mathbf{w_i}$ is a weight vector that represents the real-valued relative importance of each component of the information $\mathbf{x}$ for deciding upon choice i. Hence, one may view

$$G_i(\mathbf{x}) = \mathbf{w}_i \cdot \mathbf{x} \tag{3}$$

as representing the degree to which the available information $\mathbf{x}$ favors a decision of i to the agent. The β parameter determines the degree of stochasticity, so that probability distribution arising from the model depends more strongly on the G_i for large values of β. When $\beta = 0$, the choice is made uniformly at random, independently of the G_i, and in the limit where $\beta \to \infty$, the option i' corresponding to the greatest value of G_i is chosen deterministically, with ties being broken randomly.

To allow for biases in an agent's decision, we introduce into 2 a bias term b_i, which we add to G_i. To retain the vector notation for G_i in 2, we put the bias term as the first element of the weight vector, $w_i^{(0)} = b_i$, and let the corresponding element of the input vector be unity $x^{(0)} = 1$.

We denote the possible choice in a given round of the PDG c, for 'cooperate', and d for 'defect'. The probability distribution over the choices is is then given by:

$$\begin{aligned} p_c &= \frac{\mathrm{e}^{\beta \mathbf{w}_c \cdot \mathbf{x}}}{\mathrm{e}^{\beta \mathbf{w}_c \cdot \mathbf{x}} + \mathrm{e}^{\beta \mathbf{w}_d \cdot \mathbf{x}}}, \\ p_d &= \frac{\mathrm{e}^{\beta \mathbf{w}_d \cdot \mathbf{x}}}{\mathrm{e}^{\beta \mathbf{w}_c \cdot \mathbf{x}} + \mathrm{e}^{\beta \mathbf{w}_d \cdot \mathbf{x}}}. \end{aligned} \quad h \tag{4}$$

Due to normalization, this may be simplified as

$$\begin{aligned} p_d &= \frac{1}{1 + \mathrm{e}^{-\beta \mathbf{w} \cdot \mathbf{x}}}, \\ p_c &= 1 - p_d, \end{aligned} \tag{5}$$

where

$$\mathbf{w} = \mathbf{w}_d - \mathbf{w}_c \tag{6}$$

Based on this, we tried fitting several different models to the data:

1. One model in which the state vector $\mathbf{x}$ of eq. 3 contains only information on defecting behavior, i.e. indicator variables for whether the player and their opponent defected in the previous round, and the payouts obtained by the player and the random neighbor, in the event they defected. For instance, the indicator variable denoting whether the player defected in the previous round would be $\delta_{p,d}$, i.e. a Kronecker delta taking the value 1 if the player defected in the previous round and 0 otherwise, and the payout variable would be $\delta_{p,d} \cdot f_p$, i.e. the player's payout from the previous round if they defected, and zero otherwise.
2. A similar model, but allowing separate values and indicator variables for cooperation.
3. A model as the above, but allowing the bias term to depend on the previous action taken by the player and their opponent.

The free parameters of each model were then fitted using the COBYLA optimization method[33] to minimize the negative log-likelihood of model. Some constraints were imposed upon the parameters to avoid performance degeneracies in the parameter space - β was constrained to positive values, and the norm $|\mathbf{w}|$ of the weight vector was set to unity. The fitting procedure was repeated 10 times with parameter vectors randomly initialized in each run to mitigate the problem of local optima. The majority of runs converged and resulted in very similar negative log-likelihoods, with a few outliers at greater values, confirming the necessity of multiple runs of the fitting procedure. For the majority of runs which both converged and had similar likelihoods, the parameter vectors returned by the algorithm were closely clustered together. This was not the case when the aforementioned constraints were omitted, indicating that the constraints were indeed necessary to remove parameter space degeneracies. Performance metrics for the best fits for each model are summarized in table 2.

In addition to the models described above, two additional types of models were considered. One such type of models was similar to models 2 and 3 above, but instead of incorporating a bias term in 3 the bias would be outside the exponential function so $p_d \propto \alpha\, \mathrm{e}^{\beta \mathbf{w}_d \cdot \mathbf{x}}$, turning 5 into $\frac{1}{1+C\,\mathrm{e}^{-\beta \mathbf{w}\cdot\mathbf{x}}}$, where $C = \frac{1-\alpha}{\alpha}$. However, this resulted in the same values of the performance metrics as models 2 and 3. We also investigated models which took history from the previous two rounds, rather than just one, for the participants into account. This, however, resulted in a slightly worse fit to data, as well as higher model complexities. In addition, we tried a model that also explicitly took into account the T parameter from the payout matrix. This slightly increased model likelihood but decreased the AIC score due to the additional model complexity. For this reason, we proceed with model 3, without the T parameter.

The state vector $\mathbf{x}$ consists of the following components: An indicator of the immediate game history H available to the player, i.e. of whether they and their opponent cooperated or defected in the previous round, as well as the payouts they, and a random person in their neighborhood, received from cooperating and defecting in the previous round. $\mathbf{x}$ may be written as

$$\mathbf{x} = \left(\delta_{H,cc}, \delta_{H,cd}, \delta_{H,dc}, \delta_{H,dd}, \delta_{s,d}p_s, \delta_{s,c}p_s, \delta_{n,d}p_n, \delta_{n,c}p_n\right)', \tag{7}$$

where $\delta_{H,ij}$ is a Kronecker delta which is 1 if the player and their opponent played strategies i and

Model	n	$-\ln \mathcal{L}(\mathbf{w})$	Accuracy	AIC	$F1$
1	6	3011	.76	6059	.83
2	10	2777	.81	5614	.87
3	9	2725	.86	5503	.90

Table 2. Summary of various performance metrics for the models. The table displays the number of model parameters n, and the negative log-likelihood $-\ln \mathcal{L}(\mathbf{w})$ of each model along with its accuracy. As more complex models would be expected to fit any data better, we also provide the Akaike Information Criterion (AIC) score[34]. As the data are unbalanced (with many more choice to defect than to cooperate), we also provide the $F1$ score for the models.

j, respectively, in the previous round. Similarly, s and n are used as indices of the player themselves and the random neighbor, respectively, with p denoting payout.

Parameter adjustments for simulations

To account for the fact that most simulation approaches use a weak PDG, the model requires a few adjustments after the fitting procedure. First, the fitted heuristics display a very strong bias towards defection, possibly because participants in the experiment played a strong prisoner's dilemma game, with a game matrix given by $T = t, R = 2, P = 1, S = 0$, with t lying in $(2, 4)$, whereas in our simulations, in order to align with the literature, we use $T = t, R = 1, P = S = 0$, with t in $[1, 2]$. Second, the stochasticity in the fitted model leads to low stability in clusters of similarly acting agents, and negates interesting network effects[35]. Third, in accordance with other literature finding that experimentally determined neighbor influence is quite low[36], which adversely affects simulations[35].

The latter obstacle we overcome simply by fixing the weights representing the impact from neighbor payouts on an individual's choice to the same as the weights for the individual's own payouts. This matches well with the literature, in which every individual heuristic we encountered also treated payouts for the individual in question and their neighbors on equal footing. The problem of stability we mitigated by enforcing a rule that if an individual seeking to update their strategy, and the randomly selected individual neighbor with whom they compared strategies and payouts, had both followed the same strategy, the agent would deterministically choose that strategy. Finally, to compensate for the increased incentive to defect in the strong vs. weak prisoner's dilemma game, we shifted the bias terms so biases for the neutral data were at zero, while retaining differences between bias terms for the three terminologies. This adjustment was performed in the following way: From β and the weight vectors w (which include the bias terms) in eq. (5), a vector of 'absolute weights' $\mathbf{V}$ are computed as $\mathbf{V} = \beta \cdot \mathbf{w}$. These are equivalent to the weights in e.g. a normal logistic regression model. Notice that, as we've used the constraint $\mathbf{w}$ is L^2-normalized, we have $\beta = |\mathbf{V}|$

For a given terminology exposure i, the corresponding vector $\mathbf{V}_i$ may be thought of as a concatenation of the weight vector $\mathbf{v}_i$ and a vector of biases $\mathbf{b}_i$, i.e. $\mathbf{V}_i = \mathbf{v}_i \oplus \mathbf{b}_i$. This vector is then offset by the biases from the neutral terminology, i.e.

$$\mathbf{V}'_i \leftarrow \mathbf{V}_i - \mathbf{0} \oplus \mathbf{b}_n, \tag{8}$$

	Adjusted			Raw fits		
	C	N	I	C	N	I
w_{sd}	.23	.26	.22	.31	.47	.34
w_{sc}	.21	.24	.17	.28	.36	.31
w_{nd}	.23	.26	.22	.017	-.01	.015
w_{nc}	.21	.24	.17	-.042	-.03	-.035
b_{cc}	-.11	0	.054	-.81	-.78	-.75
b_{cd}	-.049	0	.1	-.42	-.42	-.25
b_{dc}	-.033	0	.054	.00	.052	.17
b_{dd}	-.046	0	-.0072	-.02	.047	.037
β	4.6	3.9	5.7	3.5	3.0	2.7

Table 3. Parameter values for the agent logit model. The rightmost columns contain the values obtained directly by fitting to experimental data, and the leftmost columns show the adjusted values.

where $\mathbf{b}_n$ represents the biases in the neutral model. We may rewrite this as $\mathbf{V}' = \beta'\mathbf{w}'$ for consistency with previous notation. The parameters for the three models (C, N, and I) after this transformation are displayed in table 3

The models given by the parameters in table 3 are visualized in fig. 5.

In the main paper, we investigate the effects of terminologies by running simulations in which varying fractions of the individual agents employ the decision heuristics based on the collectivist and individualist terminologies, respectively. We probe this through the parameter ρ_I, which denotes the fraction of agents that are randomly assigned to follow the model based on the individualist terminology, labeled 'I' in fig. 5, whereas the remaining $1 - \rho_I$ follow the model based on the collectivist terminology. Hence, values of $\rho_I = 0$ and $\rho_I = 1$ correspond to 'pure' systems in which every agent employs the heuristics from the collectivist and individualist terminologies, respectively, whereas intermediate values correspond to 'mixed' systems in which both groups of agent coexist.

We consider the interactions between terminologies, as expressed by ρ_I, and the 'temptation to defect' parameter t, and a range of quantities, such as the fraction of agents defecting, the mean payouts for all nodes in the network, a 'pairing measure' capturing to which degree cooperating agents are connected to fellow cooperators at a disproportional rate, etc. As agents embedded in social networks are known to exhibit a high degree of homophily in terms of communication and media consumption[37, 38, 39], we also investigate the effects of increasing the terminology homophily by assigning terminologies in way which makes agents exposed to similar terminologies more likely to be connected.

We present a brief overview and explanation of these quantities, and give summarize exploratory analyses of their interplay with networks structure and clustering in the SI.

Acknowledgments

The authors are grateful to Jason McKenzie Alexander for his insightful comments, to the employees and volunteers at the behavioural research lab at LSE, and to Erik Mohlin, Marco Islam, and Alexandros Rigos for helpful discussions.

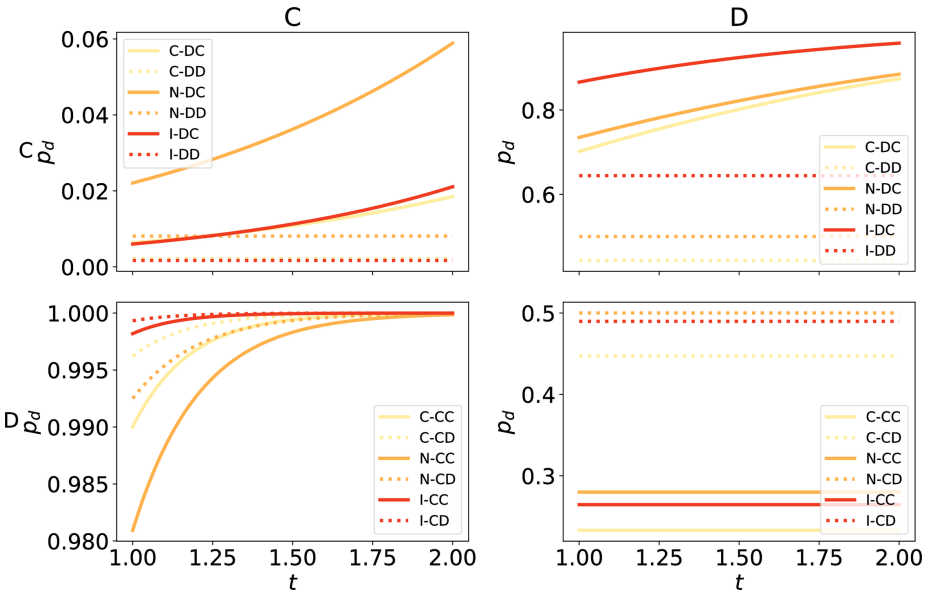


Fig 5. Visual depiction of the agent heuristics with parameters displayed in table 3. Each plot corresponds to the outcome of a round based on which a given agent is making a decision to potentially update their strategy. Rows correspond to the agent's previous strategy, and columns to their opponent's strategy, so each plot corresponds to the corresponding cell in the game matrix. Each line corresponds to a terminology (C, N, or I), and the strategies chosen by the agent's neighbor, and the neighbor's opponent in the previous round. For example, the label "N-DC" corresponds to a model based on the neutral terminology, and the situation where the agent's neighbor defected while their opponent cooperated. The x-axis represents the t parameter, and the y-axis the probability of the agent defecting p_d. Note that we do not show lines for situations where the agent and neighbor played the same strategy, as the agent retains their strategy in that case.

References

1. Tversky, A. & Kahneman, D. The framing of decisions and the psychology of choice. *science* **211**, 453–458 (1981).

2. Molden, D. C. Understanding priming effects in social psychology: What is "social priming" and how does it occur? *Social Cognition* **32**, 1–11 (2014).

3. Marwell, G. & Ames, R. E. Economists free ride, does anyone else. *Journal of public economics* **15**, 295–310 (1981).

4. Gerlach, P. The games economists play: Why economics students behave more selfishly than other students. *PloS one* **12**, e0183814 (2017).

5. Bicchieri, C. *The grammar of society: The nature and dynamics of social norms* (Cambridge University Press, 2005).

6. Szabó, G. & Fáth, G. Evolutionary games on graphs. *Physics Reports* **446**, 97–216 (2007).

7. Lieberman, E., Hauert, C. & Nowak, M. A. Evolutionary dynamics on graphs. *Nature* **433**, 312–316 (2005).

8. Davies, S. R. Constructing communication: Talking to scientists about talking to the public. *Science Communication* **29**, 413–434 (2008).

9. Sugimoto, C. R. & Thelwall, M. Scholars on soap boxes: Science communication and dissemination in ted videos. *Journal of the American Society for Information Science and Technology* **64**, 663–674 (2013).

10. Kirby, D. A. The changing popular images of science. *The Oxford Handbook of the Science of Science Communication* 291–300 (2017).

11. Carter, J. R. & Irons, M. D. Are economists different, and if so, why? *Journal of Economic Perspectives* **5**, 171–177 (1991).

12. Frank, R. H., Gilovich, T. & Regan, D. T. Does studying economics inhibit cooperation? *Journal of economic perspectives* **7**, 159–171 (1993).

13. Ifcher, J. & Zarghamee, H. The rapid evolution of homo economicus: Brief exposure to neoclassical assumptions increases self-interested behavior. *Journal of Behavioral and Experimental Economics* **75**, 55–65 (2018).

14. Ferraro, F., Pfeffer, J. & Sutton, R. I. Economics Language and Assumptions: How Theories can Become Self-Fulfilling. *Academy of Management Review* **30**, 8–24 (2005).

15. Cappelen, A. W., Nygaard, K., Sørensen, E. Ø. & Tungodden, B. Social preferences in the lab: A comparison of students and a representative population. *The Scandinavian Journal of Economics* **117**, 1306–1326 (2015).

16. Mas-Colell, A., Whinston, M. D., Green, J. R. *et al. Microeconomic theory*, vol. 1 (Oxford university press New York, 1995).

17. Binmore, K. Just playing: Game theory and the social contract II (1997).

18. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science (New York, N.Y.)* **211**, 1390–6 (1981).

19. Milinski, M. TIT FOR TAT in sticklebacks and the evolution of cooperation. *Nature* **325**, 433–435 (1987).

20. Trivers, R. L. The evolution of reciprocal altruism. *The Quarterly review of biology* **46**, 35–57 (1971).

21. Nowak, M. A. & May, R. M. Evolutionary games and spatial chaos. *Nature* **359**, 826–829 (1992).

22. Blume, L. E. The Statistical Mechanics of Strategic Interaction. *Games and Economic Behavior* **5**, 387–424 (1993).

23. Rand, D. G. & Nowak, M. A. Human cooperation. *Trends in Cognitive Sciences* **17**, 413–425 (2013).

24. Santos, F. C. & Pacheco, J. M. Scale-Free Networks Provide a Unifying Framework for the Emergence of Cooperation. *Physical Review Letters* **95**, 098104 (2005).

25. Nowak, M. A. Five rules for the evolution of cooperation. *science* **314**, 1560–1563 (2006).

26. Stopczynski, A. *et al.* Measuring large-scale social networks with high resolution. *PloS one* **9**, e95978 (2014).

27. Conover, W. J. & Iman, R. L. On multiple-comparisons procedures. *Los Alamos Sci. Lab. Tech. Rep. LA-7677-MS* 1–14 (1979).

28. Bicchieri, C. & Xiao, E. Do the right thing: but only if others do so. *Journal of Behavioral Decision Making* **22**, 191–208 (2009).

29. Nyborg, K. *et al.* Social norms as solutions. *Science* **354**, 42–43 (2016).

30. Adami, C., Schossau, J. & Hintze, A. Evolutionary game theory using agent-based methods. *Physics of life reviews* **19**, 1–26 (2016).

31. Sapiezynski, P., Stopczynski, A., Lassen, D. D. & Lehmann, S. Interaction data from the Copenhagen Networks Study. *Scientific Data* **6**, 315 (2019).

32. Wu, Z.-X., Guan, J.-Y., Xu, X.-J. & Wang, Y.-H. Evolutionary prisoner's dilemma game on Barabási–Albert scale-free networks. *Physica A: Statistical Mechanics and its Applications* **379**, 672–680 (2007).

33. Powell, M. J. D. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in optimization and numerical analysis*, 51–67 (Springer, 1994).

34. Akaike, H. A new look at the statistical model identification. *IEEE transactions on automatic control* **19**, 716–723 (1974).

35. Gracia-Lázaro, C., Cuesta, J. A., Sánchez, A. & Moreno, Y. Human behavior in Prisoner's Dilemma experiments suppresses network reciprocity. *Scientific reports* **2**, 325 (2012).

36. Grujić, J., Fosco, C., Araujo, L., Cuesta, J. A. & Sánchez, A. Social Experiments in the Mesoscale: Humans Playing a Spatial Prisoner's Dilemma. *PLoS ONE* **5**, e13749 (2010).

37. Halberstam, Y. & Knight, B. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of Public Economics* **143**, 73–88 (2016).

38. Hermida, A., Fletcher, F., Korell, D. & Logan, D. SHARE, LIKE, RECOMMEND. *Journalism Studies* **13**, 815–824 (2012).

39. Dvir-Gvirsman, S. Media audience homophily: Partisan websites, audience identity and polarization processes. *New Media & Society* **19**, 1072–1091 (2017).

40. Barabasi, A.-L. & Albert, R. Emergence of scaling in random networks. *Science (New York, N.Y.)* **286**, 509–12 (1999).

41. Sekara, V. & Lehmann, S. The strength of friendship ties in proximity sensor data. *PLoS ONE* **9**, 1–14 (2014). `arXiv:1401.5836v3`.

Self-interested behaviour as a social norm

Supplementary Information Appendix

Kamilla Buchter, Bjarke Mønsted, & Sune Lehmann

Experimental setup

Design of the first experiment

The experiment had four parts. First, participants were presented with the title "A Choice Experiment" and read a general description of the experiment. Second, they were asked to read a description of the game and answer five control questions which ensured that they understood the rules of the game. Participants could not proceed to the game before all five control questions were answered correctly. Third, the participants played ten rounds of PDG. After each round, participants were informed about the choice made by their opponent, their own pay-off from the game, and the choice and pay-off of another random participant.[1] The participants were then asked a follow-up question about the random participant in order to make them engage with this information. The participants could not continue to the next round of PDG before they had answered the follow-up question correctly. Finally, after playing the ten rounds, participants were asked to state whether they had played this type of game before, and if they had guessed the hypothesis tested in the study. The experiment took less than 15 minutes to complete.

The PDG played by the participants had the pay-off structure $T > R > P > S$ where $S = 0$, $P = 1$, $R = 2$, and T was selected uniformly at random from the interval $(2, 4)$. The two strategies were called *cooperate* and *defect*. The pay-off structure is depicted in table 6. The entire experimental set-up can be seen at: http://ahura.herokuapp.com/.[2]

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	(2,2)	(0,T)
	Defect	(T,0)	(1,1)

Fig 6. Structure of the PDG played in the experiment. The numbers indicate the points that players can win, where $T \in (2, 4)$ is chosen at random for each player. The value of T for each participant remains the same throughout the experiment.

To test the effect of microeconomics textbooks terminology, participants were randomly allocated to one of three categories in the experiment. The control category used a neutral terminology and did not introduce a microeconomics concept. The second category used an individualist terminology and asked participants to read a text excerpt stating that in game theory the word *rational* is used to denote the strategy of defecting. The third category used a collectivist terminology and asked participants to read a text excerpt stating that in game theory the word *optimal* is used to denote the strategy of cooperation. The two text excerpts were designed to be similar in their formulation, so that the only difference is whether they apply a positively laden word to the strategy of defection

[1] The information about the random participant was generated by the computer at random.

[2] To go to the experiment, enter a random sign in the field for an identification code, press "I agree - continue to study", and press "submit".

or to the strategy of cooperation. Both words relate to concepts used in microeconomics, though rational behaviour is more widely used than socially optimal outcomes.[3] Next, we provide a detailed account of the stimuli used in each of the three categories.

0.0.1. Neutral terminology

The first category in the experiment used a *neutral terminology* in order to provide a control version to which the other two terminologies can be compared. In this category, participants were greeted with the sentence "Welcome! You are about to take part in a study on how we make strategic choices." Further, *the other participant* was used to describe other players in the experiment. The category did not introduce a microeconomics concept and the control questions pointed both to dominant strategies and to the benefit of cooperation without any normative wording:

1. If, in a given round, the other participant and you both play 'cooperate', how many points do you receive?
2. In a given round, you choose 'defect' and receive 1 point. Which strategy did the other participant choose?
3. If the other participant plays 'defect' in a given round, which strategy should you choose to ensure that you get the greatest possible number of points?
4. If the other participant plays 'cooperate' in a given round, which strategy should you choose to ensure that the two of you receive the greatest possible total number of points? Hint: for each strategy combination, sum the payoffs you and the other participant will receive.
5. Assume that in a given round you choose 'defect' and receive T. Which strategy did the other participant choose?

Finally the follow-up question after each round said:

- A random participant played [cooperate/defect] in the previous round, and received a payout of [T/2/1/0].
  Which strategy did that player's opponent choose?

This ensured that participants engaged with the information they received about the random participant's game.

0.0.2. Individualist terminology

The second category of the experiment used an *individualist terminology* to mirror the language of standard microeconomics textbooks. The terminology was introduced through four changes to the experiment.

[3]Notice that the texts presented in the experiment provide a somewhat simplified version of game theory since rationality and social optimality is concerned with utilities rather than monetary gains. Thus, the concepts only apply to the experiment under the assumption that participants will increase their utility by increasing their wealth.

First, the title of the experiment was accompanied by a small subtitle "A study on rationality" and participants were greeted with the sentence "Welcome! You are about to take part in a study on rationality." Further, *your opponent* was used to describe the other players. These situational cues were supposed to indicate to the participants that they were in a competitive situation.

Second, participants were asked to read a short text introducing the microeconomics concept of *rational* to describe the choice of defecting in the game. The text was:

> A concept of particular interest in this study is the notion of **rationality**. In game theory, we say that it is **rational** for a player to choose a strategy, if the strategy is guaranteed to result in a greater payoff to the player, regardless of which strategy their opponent plays. Conversely, we say that it is **irrational** for a player to choose a strategy that does not guarantee the highest possible payoff (regardless of what the other player chooses), if a strategy that does so is available. The following contains a few control questions to ensure that you understand these concepts and their relation to the game.

Third, control questions 3-5 were changed in order to ensure that participants understood the concept of *rationality* and knew how to apply it. The three control questions were:

3. If your opponent plays 'cooperate' in a given round, which strategy should you choose to ensure that you get the greatest possible number of points?
4. If your opponent plays 'defect' in a given round, which strategy should you choose to ensure that you get the greatest possible number of points?
5. Given your answers to the above, how would the 'defect' strategy be classified according to game theory?

Finally, the follow-up question after each round of the game was changed:

- A random participant played [cooperate/defect] in the previous round, and received a payout of $[T/2/1/0]$.
  How does game theory categorize this strategy?

The participants could either answer *rational* or *irrational*. This change ensured that participants engaged with the information about the random participant and that they used the microeconomics terminology throughout the experiment.

0.0.3. Collectivist terminology

The third category used a *collectivist terminology*. The collectivist terminology was designed to mirror the individualist terminology by having a parallel sentence structure. The terminology was introduced through four changes to the experiment.

First, the title of the experiment was accompanied by a small subtitle "A study on cooperation" and participants were greeted with the sentence "Welcome! You are about to take part in a study on cooperation." Further, *your co-player* was used to describe the other participants in the experiment.

These changes were intended to provide the participants with a situational cue that they were in a cooperative situation.

Second, participants were asked to read a short text introducing the concept of *optimal* to describe the choice of cooperating in the game:

> A concept of particular interest in this study is the notion of **social optimality**. In game theory, we say that an outcome is socially optimal if it results in the largest overall payoff and if no one can be made better of without making someone else worse off. We call a strategy that can lead to a socially optimal outcome **optimal**. Conversely, we call a strategy **suboptimal** if it cannot lead to a socially optimal outcome. The following contains a few control questions to ensure that you understand these concepts and their relation to the game.

Third, control questions 3-5 were changed to ensure that participants understood the concept of optimality and knew how to apply it:

3. If your co-player plays 'cooperate' in a given round, which strategy should you choose to ensure that the two of you receive the greatest possible total number of points (i.e. which choice maximizes the sum of the points that you and your co-player receive)?
4. If your co-player plays 'defect' in a given round, which strategy should you choose to ensure that the two of you receive the greatest possible total number of points? (i.e. which choice maximizes the sum of the points that you and your co-player receive?)
5. Given the above how may we classify the role of the 'cooperate' strategy in increasing overall wealth?

Finally, the follow-up question was changed to:

- A random participant played [cooperate/defect] in the previous round, and received a payout of $[T/2/1/0]$.
  Which of the following best describes this strategy choice?

The participants could either answer *optimal* or *suboptimal*.

Design of the second experiment

In order to ensure that the situation in the second experiment is comparable to the situation in the first experiment, we used the same experimental design as reported above. However, we made one change to the experiment. In each round of the PDG, participants were asked two additional questions, before they indicated which strategy they wanted to play. The first question was designed to ask about the participants' empirical expectations, while the second question was designed to ask about participants' normative expectations. The questions were:

- Which strategy do you think the other participant will choose?

- We also ask the other participant which strategy they think **you** will choose. What do you think the other participant answers?

The participants could answer *cooperate* or *defect* to each question. For participants in the collectivist category in the experiment, "the other participant" was changed to "your co-player". For participants in the individualist category, it was changed to "your opponent".

Interplay between heuristics and network

This section provides an overview of the classes of real and artificial networks we considered for analyses, as well as a range of possible update heuristics for the agents embedded in the simulations. As simulations behave differently in each of the relatively large number of combinations of network types and update heuristics, we provide an overview here along with some qualitative reasons for our choice of focus in the main paper.

The networks considered fall in one of two categories - real, and artificial, i.e. constructed using real-world data, and constructed computationally, starting from a small set of simple rules. The artificial networks under consideration are:

- A simple 2-dimensional square lattice (SL), such as the one considered in Nowak and May's famous 1992 paper on evolutionary games[21]. In this network, each node is connected only to its 4 neighbours - north, south, east, and west, with periodic boundary conditions.
- An Erdos–Rényi (ER), in which every pair of nodes (u, v) are randomly connected, each with an independent probability chosen as 1%.
- A Barási-Albert (BA) scale-free network[40] constructed by starting with a set number m of interconnected nodes, and then grown using a preferential attachment scheme in which each new node is attached to m existing nodes with probabilities proportional to the degrees of the nodes. In the literature, we encountered results from simulations on BA networks with parameter choices ranging from $m = 4$[24] to to $m = 8$[32], leading us to use $m = 6$.

All networks mentioned in the above were constructed with a size of $n = 625$ nodes.

In addition to this, a range of networks constructed from real-world data were also employed. The data used to construct all such networks comes from the Sensible DTU experiment[26] at the Technical University of Denmark. The experiment consisted of a large number (> 700) of Danish university students, who received smartphones which, with their consent, registered information regarding contact patterns, sensor information, etc. The data for a one-month observation period of this study are made publicly available[31]. From this, the following networks were constructed.

- A text message (SMS) network with any two nodes (u, v) connected if either had texted the other during a one-month observation period, containing a total of $n = 457$ nodes.

- A Facebook (FB) network consisting of $n = 800$, with users linked if they were friends on Facebook.
- A Bluetooth (BT) network with $n = 542$ nodes. This network consists of temporal 'slices' of periods of one hour. During each such time slice, an edge (u, v) is present if the corresponding users were in physical proximity of one another during the period, as identified by the Bluetooth sensors in their phones. Proximity was then detected by thresholding signal strengths to RSSI values above $-90dB$, corresponding to distances of a few metres[41].

In addition, we considered a range of different update heuristics for agents engaged in repeated games on various graphs. A brief overview of the heuristics considered, as well as some descriptions of their qualitative differences, is provided below.

1. A 'local maximum' heuristic in which agents consider themselves and their neighbourhood, and copy the strategy of whichever node received the greatest mean payout in the previous round.
2. An 'individual max' heuristic, in which agents follow the same procedure as above, but only compare themselves with a single node from their neighbourhood, which they choose uniformly at random.
3. An 'local softmax' heuristic, in which nodes consider the payouts earned by themselves in the previous round, and the average payout of neighbors using the opposite strategy. If all neighbors used the same strategy as the nodes, its strategy will not update. Otherwise, it will use the two payouts as inputs to a softmax function which determines the probabilities of the strategies.
4. An 'individual softmax' heuristic similar to 3 but comparing the node to a randomly selected neighbor.
5. A 'local stochastic softmax' heuristic, similar to 3, but without the constraint that nodes must deterministically reuse their previous strategy if nobody in their neighbourhood played the opposite strategy in the previous round.
6. A 'individual stochastic softmax' heuristic, similar to 5, but considering only a randomly selected neighbor.

The heuristics outlined in items 1 to 6 lead to different global dynamics on different networks. In figs. 7 to 12 these dynamics are shown for a range of network structures. In each figure, each row of subfigures corresponds to a network structure. The right column of subfigures summarise the influence of the 'defection temptation' parameter t. The subfigures show the fraction of nodes defecting after 10^4 simulation steps, averaged over an additional 10^3 steps (red) as well as the pairing measure (gray) obtained in a similar way. The dashed lines show the value for t resulting in the value of ρ_I that is as close as possible to the midpoint between the maximum and minimum values for ρ_I.

For this value t^*, an additional simulation was run for each network. The end states after 10^4 time steps are illustrated in the left columns of subfigures, with red and blue denoting defectors and cooperators, respectively. The central columns show the degree distributions of the networks, coloured based on defection rates - for each degree, a score was computed by taking the average number of times nodes of each degree defected in the last 10^3 iterations of the simulation. The nodes were then coloured based on the ranks of those scores, with more defection/cooperation corresponding to more blue/red colours.

The above preliminary investigations reproduce several findings from the literature. One example is the sharp phase transition on the square lattice in figure 7, first observed by Nowak and May. Note that the transition occurs at slightly lower values of t, because we opted for an asynchronous update scheme in which 10% of nodes, rather than all of them, changed their strategy in each round. Other such findings include under stochastic heuristics, network structure has little to no effect[35], as shown in e.g. fig. 11. Finally, we reproduce the finding that, when using a non-deterministic update heuristic, BA and ER networks facilitate similar levels of cooperation[32].

In the first experiment - from which the data we intended to fit these heuristics to originated - participants were presented with information about a single node in their vicinity, leading us to limit ourselves to the 'individual' heuristics presented above. The deterministic examples of such heuristics, such as that described in 2 are ill-suited to fit to data, a single data point can have a likelihood of zero. At the same time, we wanted a model for individual behaviour which, like those encountered in the literature, can accommodate relatively stable regions of cooperators and defectors. Hence, we ended up fitting a stochastic individual softmax model like that described in item 6, which we then adjusted to have such properties. The methods section of the main paper describes these adjustments as well as the fitting procedure.

Effects of clustering

In order to investigate the effects of distributing different update heuristics across a network in a non-uniform fashion, we devised a method of sampling from an ensemble of networks with a continuously varying degree of clustering with regards to terminology category, i.e. varying the tendency for nodes to be disproportionally connected to nodes that act according to the same model (meaning model trained on participants exposed to the collectivist, neutral, or individualist terminology).

The clustering parameter α

We did this by defining a hyperparameter α, signifying the degree of clustering. We then start with a network in which there is no category assigned to the nodes, and perform the assignment in the following fashion. A category is selected based on a predefined parameter ρ_I, so categories are chosen with $p_d = \rho_I$ and $p_c = 1 - \rho_I$. Then, with probability α, the chosen category n is assigned to a node selected using a Barabási-Albert style preferential attachment mechanism[40], in

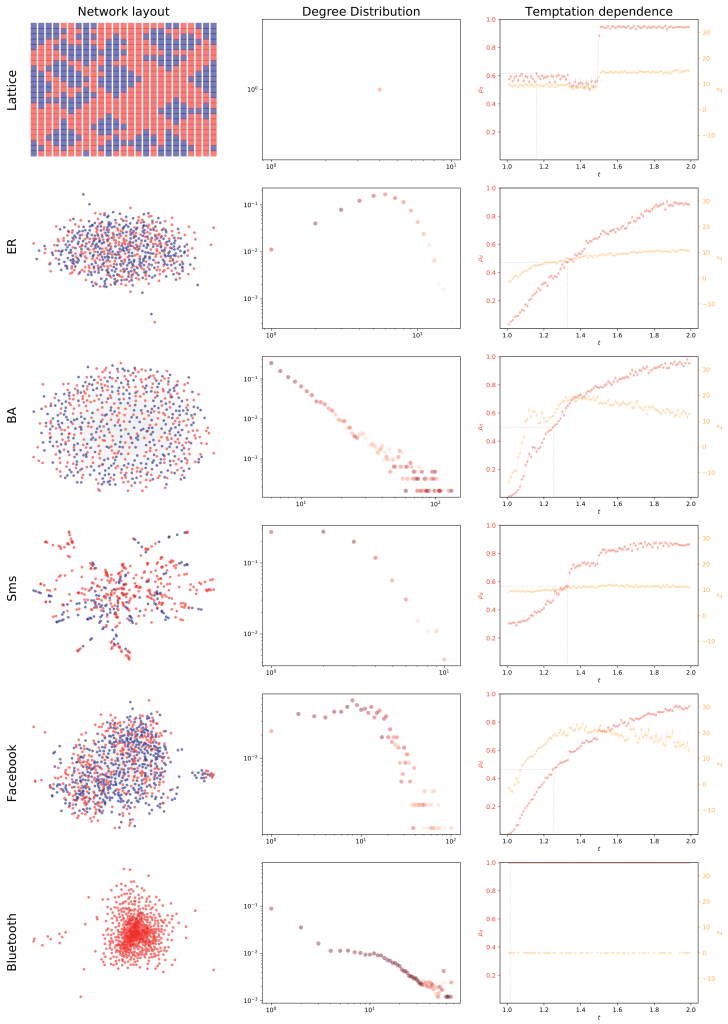


Fig 7. summary of the local maximum heuristic (1).

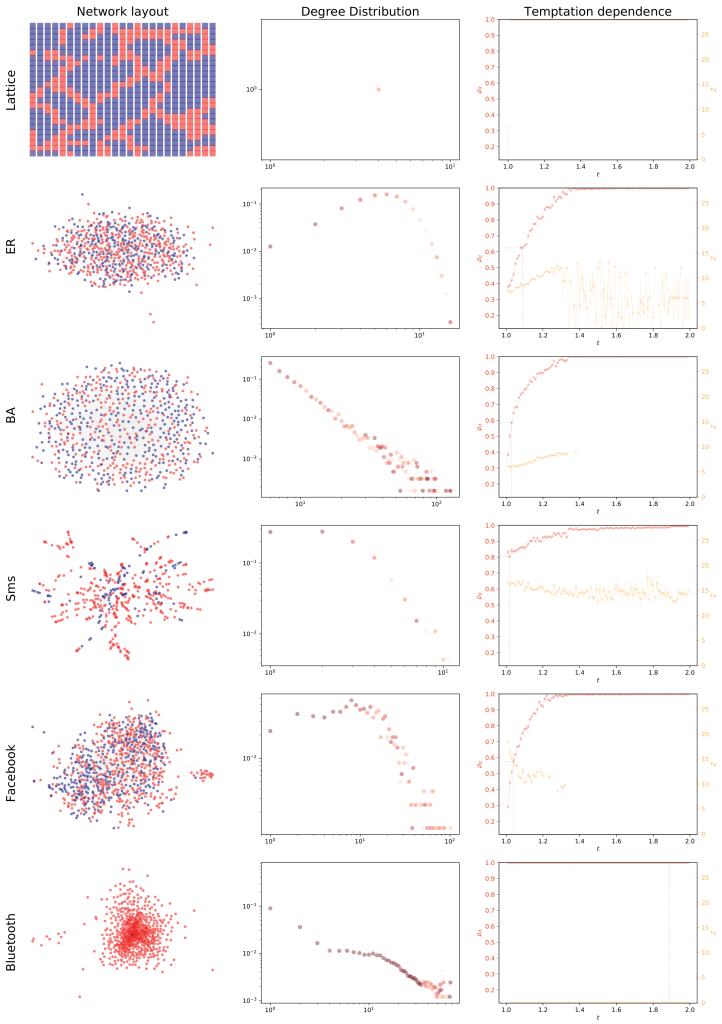


Fig 8. summary of the individual maximum heuristic (2).

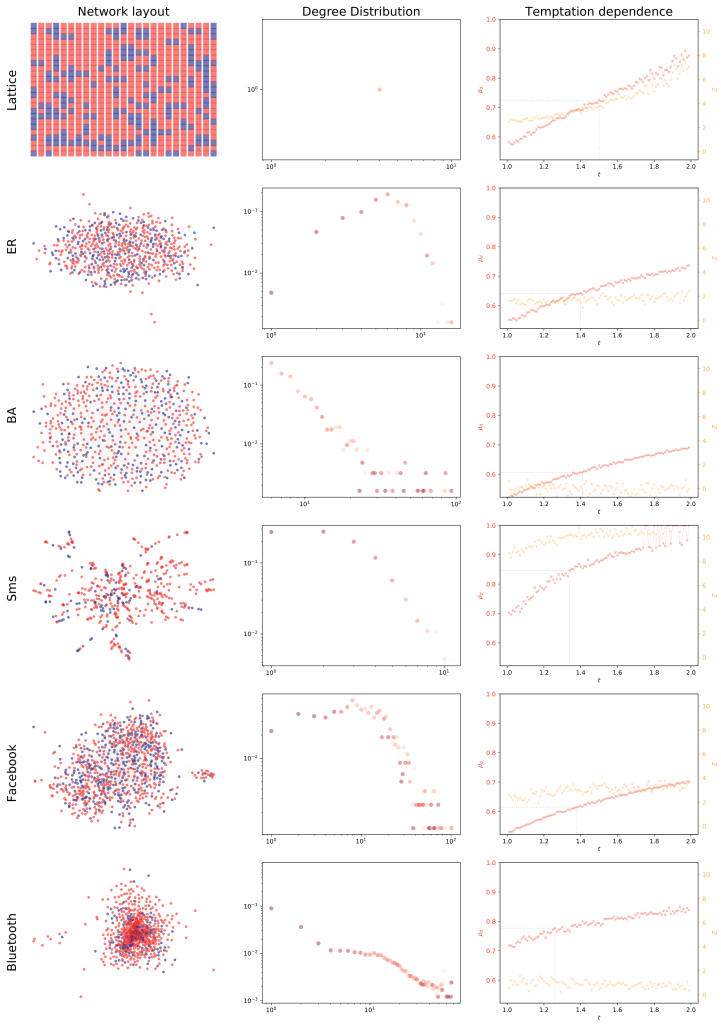


Fig 9. summary of the local softmax heuristic (3).

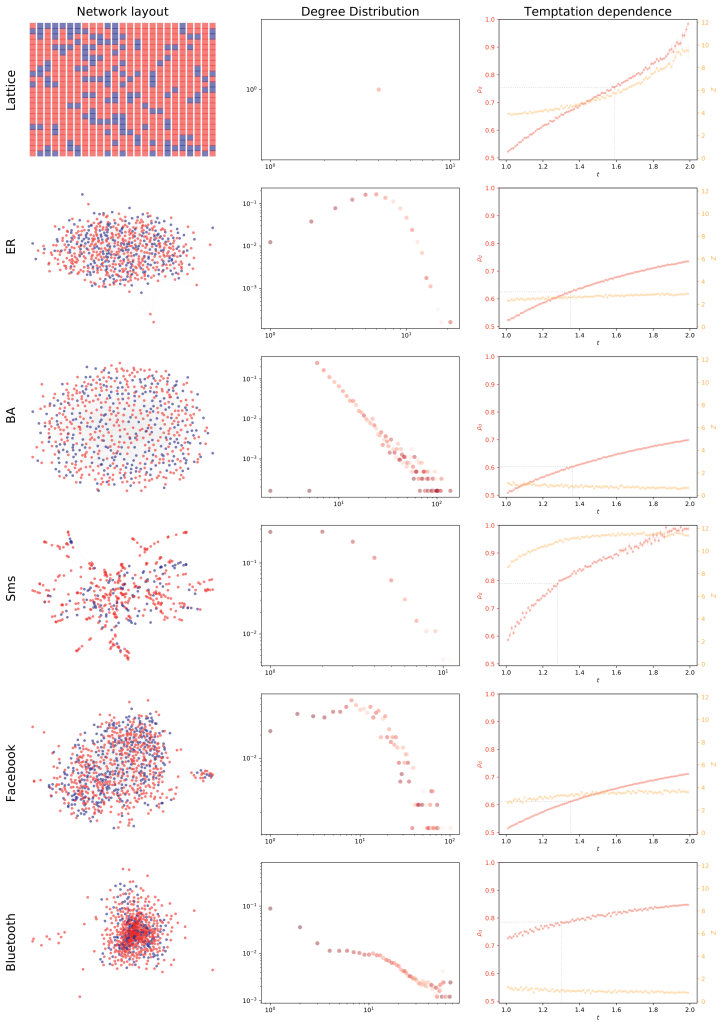


Fig 10. summary of the individual softmax heuristic (4).

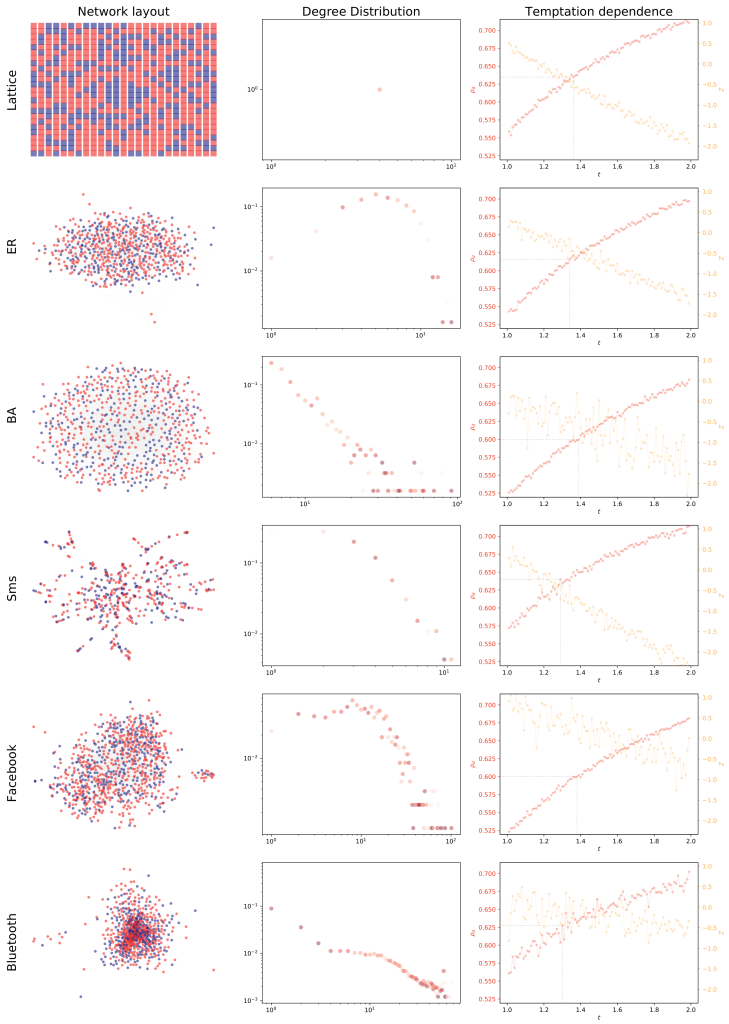


Fig 11. summary of the local stochastic softmax (5).

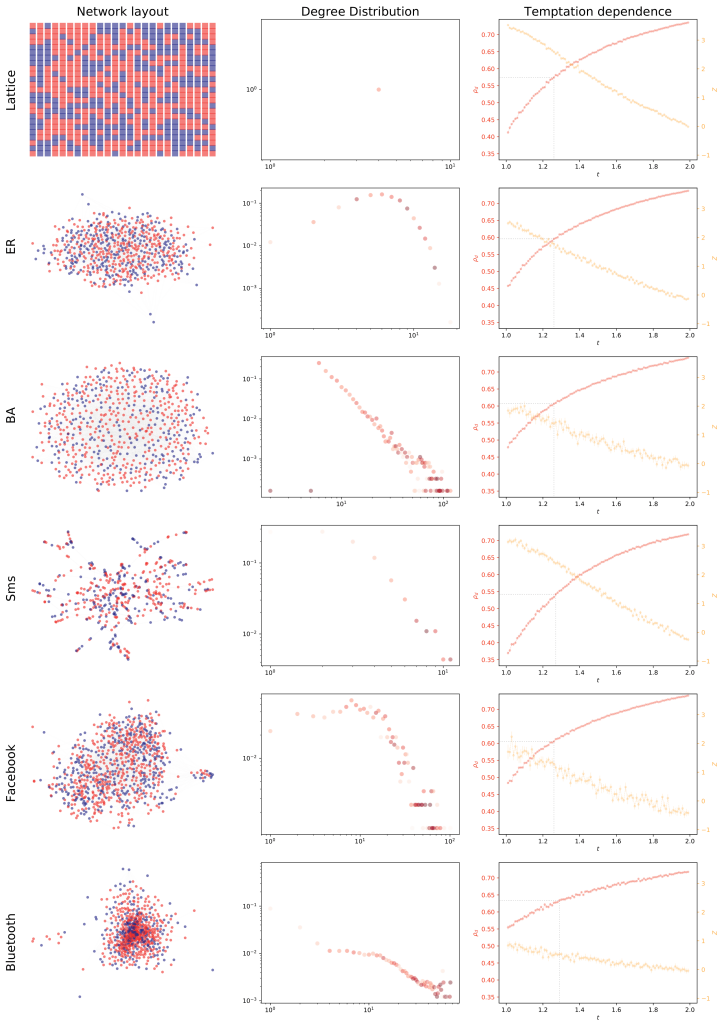


Fig 12. summary of the individual maximum heuristic (6).

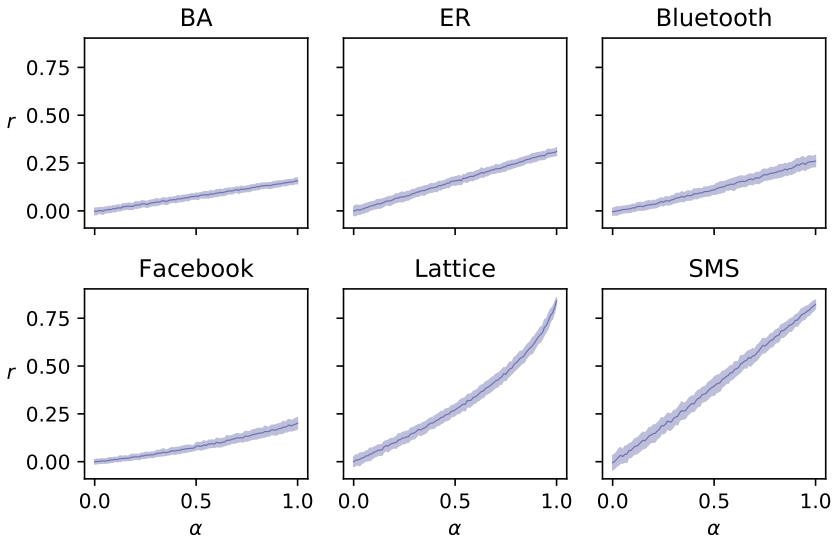


Fig 13. As the clustering hyperparameter α increases, the assortativity coefficient r of the categories grows accordingly. r appears to be growing the most in sparser networks, such as the artificial lattice, or the real text message network.

which a node u is selected with a probability proportional to the number of its neighbours having been assigned category n. With probability $1 - \alpha$, the node is selected uniformly at random. This procedure is repeated until each node has been assigned a category. Figure 13 shows the dependence of the category assortativity coefficient r on α for a range of network structures.

Simulation results

In the following, we present the results of a series of simulations and show the resulting metrics discussed in section 0.0.3. For each metric, we run a series of simulations, for a range of values of ρ_I and t, and for three values of the clustering parameter α, repeated for the SMS, FB, BA, and SL networks discussed in section 0.0.3. For each combination of network and α value, we present a 2D heatmap showing how the quantity in question changes with ρ_I and t. Each cell in these matrices is computed by running 10 simulations for 10^4 iterations, and averaging the values taken by the quantity over the last 5000 iterations. The cell is left white in cases where the measure is ill-defined - for example, the pairing measure is not defined when all nodes either cooperate or defect, as the denominator becomes zero in those cases.

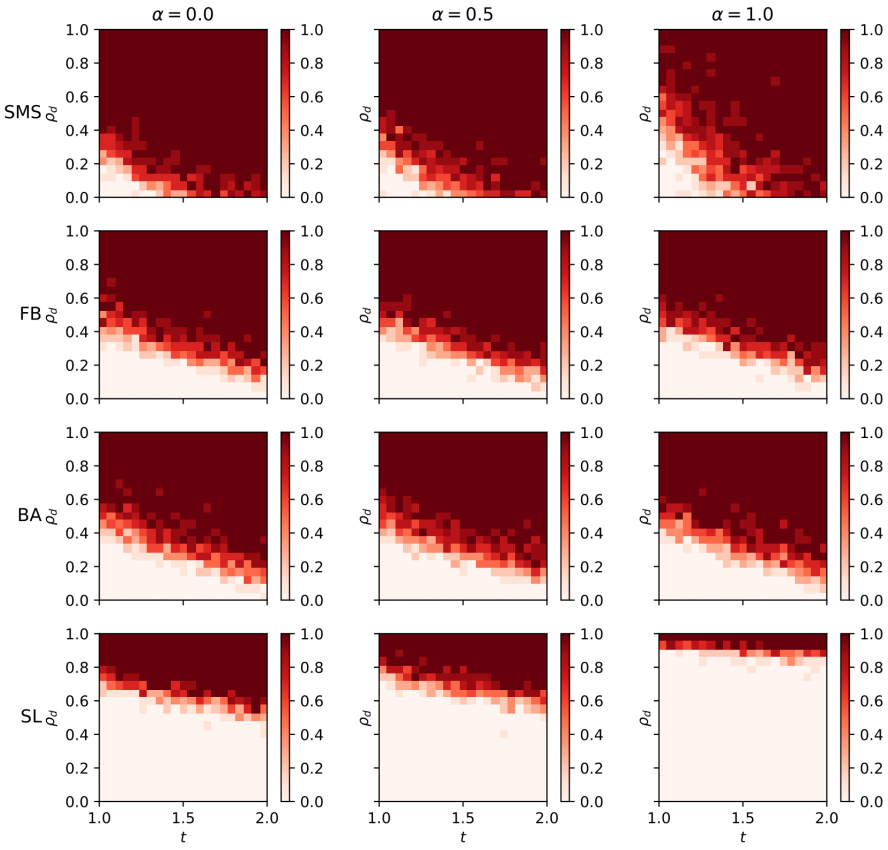


Fig 14. The fraction of defectors in the network after running the simulations.

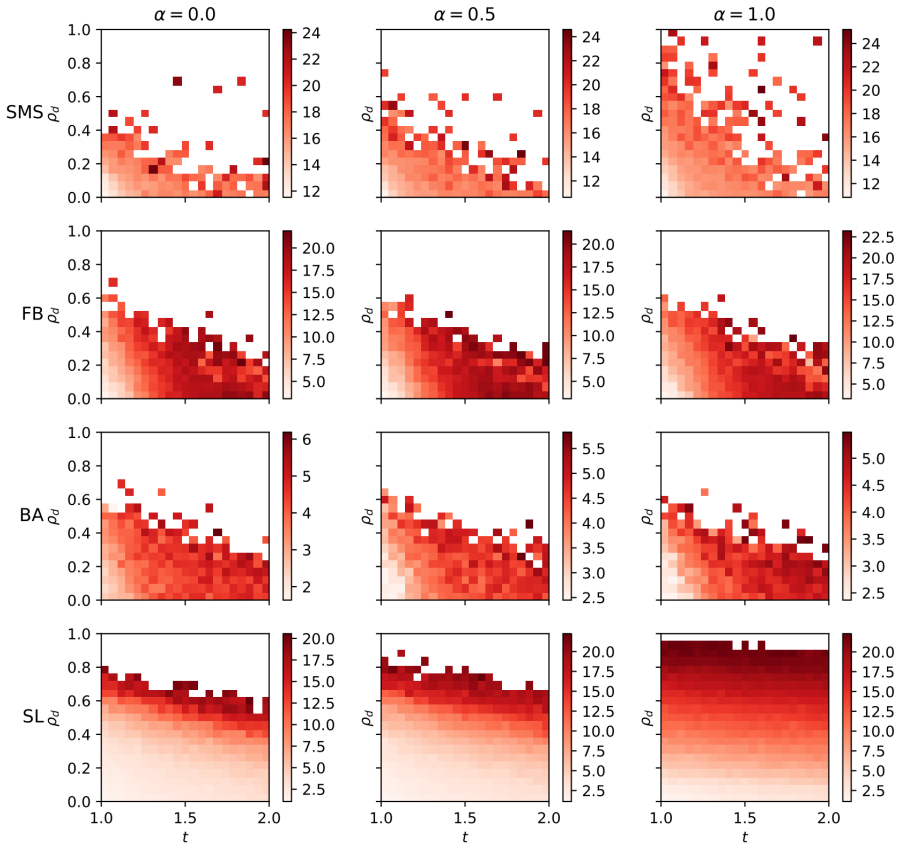


Fig 15. The prevalence measure Z_c.