Ideology, Rationality, and Revolution

An Essay on the Persistence of Oppression

Nicolas Olsson-Yaouzis

ISBN 978-91-7447-532-6

Printed in Sweden by US-AB, Stockholm 2012
Distributor: Department of Philosophy, Stockholm University

To Lovisa Olsson and the memory of Evriviades Yaouzis.

Contents

Preface **vii**

1 ***Marxismus sine stercore tauri*** **9**
1.1 Explanations 14
1.2 Organisation of the book 19

2 **Oppressive social orders** **23**
2.1 Introduction 23
2.2 The mystery of oppression 24
2.3 Oppression 26
2.4 Unjust treatment 29
2.4.1 Equality of opportunity 31
2.4.2 Maximin and welfare 33
2.5 Cases of oppression 37
2.5.1 North Korea 38
2.5.2 The American dream 40
2.5.3 The gender wage gap 42
2.6 Summary 45

3 **Rational revolutionaries** **49**
3.1 Introduction 49
3.2 Definitions and assumptions 51
3.3 A free-rider problem 56
3.4 Overcoming the free-rider problem 62
3.4.1 Creating incentives 66
3.4.2 Coordinating the Stag Hunt 70
3.5 Explanation and false assumptions 77
3.6 Summary 79

4 **The explanatory importance of ideology** **81**
4.1 Introduction 81
4.2 Oppression without gunmen 83

4.2.1 Ideology 87
4.2.2 The gender wage gap 93
4.3 Explanatory importance of ideology 101
4.4 Summary 107

5 On rain dances, mechanisms, and functional explanations 109
5.1 Introduction 109
5.2 Consequence laws 111
5.3 Mechanisms in the special sciences 118
5.3.1 Mechanisms 120
5.4 A reductionist dilemma 123
5.4.1 Social forces 125
5.4.2 Reduction 128
5.5 An epistemic argument 133
5.6 Summary 135

6 Ideology demystified 137
6.1 Introduction 137
6.2 Rosen's argument 138
6.3 Memetics 143
6.4 The replicator dynamic 146
6.5 Ideology production as information cascades 154
6.6 Summary 162

7 What is to be done? 165
7.1 Introduction 165
7.2 Confirming the ideology theory of oppression 166
7.3 Policy implications 172

References clxxvii

Preface

It is a truth universally acknowledged, that it takes a village to write a dissertation. Consequently, there are a lot of people who deserve to be acknowledged. Limited cognitive capacities prevent me from remembering and mentioning everyone who has contributed. In what follows, I would like to mention those I can remember.

This dissertation springs from a cooperative project between the department of philosophy of Stockholm University and the Department of Political and International Studies at Rhodes University, South Africa. This project, *Oppression, ideology, and democracy*, was funded in part by the Swedish Research Council (VR) and Swedish International Development Cooperation Agency (SIDA). At the time, I was a hard-core proponent of rational choice theory and could not understand why anyone would abandon the assumptions of common knowledge and perfect rationality. As time went on, however, I have had reason to revise my initial beliefs. I would like to blame this in part on the two "ideologists" of the project, Tony Fluxman and Chandra Kumar. I would also like to mention Magnus Jiborn who was connected to this project. His work on evolutionary game theory and coercion has been a great source of inspiration.

Once I began to work on the dissertation, it was as part of a new project supported by the Swedish Research Council (VR): *Explanations of repression by a minority of the majority - A research program*. I wish to thank them for the generous funding.

I spent the spring of 2010 at the Centre for Philosophy of Natural and Social Sciences (CPNSS) at the London School of Economics. I would like to thank Luc Bovens for providing an invitation to the centre. Among the researchers there at the time, Jason Alexander and Rory Smead deserve a special mention for their help with my evolutionary game theoretic problems. An early version of chapter 6 was presented at the philosophy department's PhD seminar. I am grateful for the comments from the participants, especially Ittay Nissan. The visit at CPNSS was made possible by generous grants from The Swedish Foundation for International Cooperation in Research and Higher Education (STINT) and the Knut and Alice Wallenberg Foundation.

I would like to acknowledge the contribution of my supervisors Gustaf Arrhenius and Torbjörn Tännsjö. They have both been very supportive and

helpful in all matters philosophical and practical. Especially during the intense last months of the project. And although Gustaf is a Bollinger Bolshevik, I might now add him as a Facebook-friend.

I also want to acknowledge the help of Folke Tersman who I was fortunate to have as commentator on what I thought was the final draft of the dissertation. His extensive comments helped me realize that it was not.

Most chapters have, in one form or another, been presented at the practical philosophy PhD-seminar at Stockholm University. The input from the participants and commentators at these seminars has been greatly appreciated. In particular I would like to thank Henrik Ahlenius, Per Algander, Emil Andersson, Karin Enflo, Lisa Furberg, Johan Gustafsson, Mats Ingelström, Sofia Jepsson, Jens Johansson, Sandra Lindgren, Karl Pettersson, Maria Svedberg, and Attila Tanyi for their comments. I would also like to acknowledge Jonas Olson who has not only provided valuable comments, but also concert and football tickets. Finally, thank you Katharina Berndt for your valuable comments on almost every philosophy text I have written and for being the best possible person to share an office with.

Although there are many at the department of philosophy who I have benefited from, I would like to direct a special thanks to Björn Eriksson, Hans Mathlein and Karl Karlander. Björn Eriksson has patiently discussed general questions in the philosophy of the social sciences, evolutionary theory, and TV-series. Hans Mathlein has been skeptical, not only of functional explanations, but of everything. Karl Karlander has taken the time to help me with my mathematical problems (even though he considered my applied mathematical problems a bit vulgar).

Outside the department I would like to thank Robert Östling for valuable discussions on economic models and methodology. I would also like to thank Karl Wärneryd for inspiring lectures and discussions on political economy (although I suspect that he considers a dissertation on the ideology theory of oppression a sure sign that I have gone insane).

I would also like to thank Fredrik Paues and Maria Åslund for having proofread the final manuscript. I can assure you, without their help it would have been close to unreadable.

Finally, I would like to thank family and friends who have made the long and winding road bearable. Especially Maria Åslund and Theodor Yaouzis. Maria, I could not have done this without you (and to be honest, without you I would probably do engineering instead of philosophy). Although Theodor came in at the end of the project, his help should not be understated.

1. *Marxismus sine stercore tauri*

> Nature hath made men so equall, in the faculties of body, and mind; as that though there bee found one man sometimes manifestly stronger in body, or of quicker mind then another; yet when all is reckoned together, the difference between man, and man, is not so considerable, as that one man can thereupon claim to himselfe any benefit, to which another may not pretend, as well as he. For as to the strength of body, the weakest has strength enough to kill the strongest, either by secret machination, or by confederacy with others, that are in the same danger with himselfe. [Hobbes, 1996, p. 86-7]

In the light of Thomas Hobbes' observation it seems strange that, for example, the apartheid regime in South Africa survived as long as it did. The black population did, after all, vastly outnumber the oppressive white population. Similarly, it is surprising that the oppressive one-party states of the Eastern Bloc persisted for almost half a century, that the theocracy in Iran still holds power, and that Kim Jong-Il successfully managed to hand over power to his son Kim Jong-Un in North Korea. The list of oppressive regimes where a minority has oppressed or still oppresses a majority can be made long.

Moreover, if we believe that most people have a preference for being treated justly, then we can expand the list of social phenomena that needs to be explained in light of Hobbes' observation. Marxists, who believe that capitalists treat workers unjustly by exploiting them, have to explain why class structures persist. The same goes for explaining the persistence of gender hierarchies. When these phenomena are found in Western democracies such as Sweden, the UK, or the USA, we encounter an additional mystery: in these countries neither class nor gender hierarchies are legally sanctioned as, for example, class was in India and gender hierarchy still is in Saudi Arabia.

One explanation that had great influence during a large part of the 19th and 20th centuries was the *ideology theory of oppression*. According to this theory, oppressive regimes remain in power because the members of the oppressed group somehow become convinced that the regime is legitimate.

However, the ideology theory of oppression does not claim that the oppressed are masochists that enjoy being oppressed. Rather it rests on the as-

sumption that most people have a sharp eye to their own rights and interests and feel uncomfortable at the thought of being treated unjustly. Thus, the best explanation of why the oppressed do not revolt is that they do not perceive themselves as being oppressed. They may believe that the injustices they suffer are "*en route* to being amended, or that they are counterbalanced by greater benefits, or that they are inevitable, or that they are not really injustices at all." [Eagleton, 2007, p. 27] The oppressed who hold such beliefs are sometimes said to be victims of ideology or to suffer from *false consciousness* since the beliefs conceal the reality of their oppressed situation.

In 2005, *New York Times* conducted a study where they asked Americans about their attitudes and beliefs about class in America.[1] Among other things, the respondents were asked whether intergenerational social mobility had increased in the USA over the last 30 years. Only 23% answered correctly that it had decreased since 1975. If people who would like to live in a social order with a high degree of social mobility falsely believe that they do so, then it is not surprising that the lack of social mobility persist. A proponent of the ideology theory of oppression would argue that this false belief is a good example of an ideological belief that preserves an oppressive social order.

Furthermore, according to the traditional account of the ideology theory of oppression the oppressed suffer from ideological beliefs because such beliefs *serve the function* of preserving the oppressive social order. The functional claim dates back (at least) to Karl Marx's preface to *A Contribution to the Critique of Political Economy*, where he famously asserts that the superstructure of a society (i.e., its laws, norms, ideas, etc.) is determined by its base (i.e., the relations between the productive forces). The relation between superstructure and base is such that the superstructure takes the form that best serves the function of preserving the present base. Marx did, for example, believe that this could explain not only why feudal and capitalist societies have different property laws, but also why their members tend to value different things. Feudal societies tend to value family relations more than capitalist societies since these make it easier to keep power within the family and thus preserve the feudal order.

With respect to the NYT study, a proponent of the ideology theory of oppression could say that Americans believe that social mobility has increased because these beliefs are beneficial for the economic elite; or because they serve the function of preserving the unequal oppressive economic order.

Although the ideology theory of oppression might be best suited for explaining oppression in the absence of a coercive legal framework, it can also be used to explain why oppression persist under more tyrannical regimes. It

[1] See http://www.nytimes.com/pages/national/class/ accessed May 19 2011. We will return to this study in chapter 2.

can, for example, be argued that the apartheid regime survived in part because a majority of the white and non-white population were convinced that whites and non-whites were different in some essential sense and that the apartheid laws, therefore, were necessary and not really unjust.

The ideology theory of oppression is usually associated with a Marxist tradition where the goal is to explain why oppression persists in the absence of, e.g., legally sanctioned class distinctions. However, since a number of different and seemingly incompatible interpretations of ideology can be identified in Marx's writings, it is difficult to turn directly to his texts for an account of the ideology theory of oppression.[1]

Another problem with turning directly to Marx, is that his theory was intended to be scientific. That is, he believed that his theory could provide explanations and predictions in the same way as theories in the natural sciences. However, as the 20th century progressed it became evident that many of Marx's predictions failed to materialise. If Marx's theory would be judged by the same standard as the theories of the natural sciences, then it should be rejected.

The failures of Marx's theory as a scientific theory lead some 20th century Marxists to argue that Marx had gotten his own theory wrong. Contrary to what Marx believed, it should not be judged by the same standards as the natural sciences. Rather it should be seen as a theory that allows us to interpret and understand social phenomena. For example, George Lukács's [1972] theory of class consciousness provides a Hegelian interpretation of Marx. Louis Althusser's [2001] theory of ideological state apparatuses allows us to interpret capitalist society in terms of the psychoanalytical theories of Freud and Lacan. Yet another example that can be categorised as belonging to the ideology theory of oppression is Michel Foucault's [1991] theory of society as a Panopticon.

Social scientists comfortable with the interpretative tradition were drawn to the ideology theory of oppression. However, social scientists who believed that the social sciences should be judged by the same standards as the natural sciences, found the theoretical foundations of traditional Marxist theory and the ideology theory of oppression problematic. Among these were a group of philosophers and social scientists who called themselves *analytical Marxists*. They were united by their beliefs that, on the one hand, there are some fundamental insights in Marxism, e.g., historical materialism, class, and exploitation as organising categories, and the normative claim that socialism in some form is superior to existing capitalism; and, on the other hand, that the method employed by the non-scientific Marxists was lacking.[2] The analytical Marxists were committed to the same scientific method as Marx and had little

[1] See, e.g., Terry Eagleton [2007] and Michael Rosen [1996, ch. 6].

[2] See John Roemer's introduction in Roemer [1986, p. 4].

patience with the dialectic and interpretative method of their more continental minded comrades. Gerald Cohen [2000, p. xxiii], for example, claims that the dialectic form of reasoning associated with traditional Marxism "thrives only in an atmosphere of unclear thought." Cohen goes on to point out that before the term "analytical Marxism" was coined, he talked about the movement as *non-bullshit Marxism* implying that much of traditional Marxism was bullshit in the sense that it was intellectually dishonest.[1] The group's (rather confrontative) motto became *Marxismus sine stercore tauri.*[2].

Being dismissive of the Hegelian dialectics found in Lukács, the psychoanalysis of Althusser, and the post-modernism of Foucault, most analytical Marxists gave up on the ideology theory of oppression. However, since they shared the normative outlook of most other Marxists that, e.g., workers were oppressed by capitalists, they realised that they had to provide an alternative explanation, compatible with their methodological commitments, of why such oppression persists.

Most analytical Marxists believed that the best way to explain social phenomena was with the help of rational choice and game theory. That is, by assuming that agents are (more or less) rational, self-interested, correctly informed about their situation, and that all this is common knowledge.

Such theories can easily be applied to tyrannical social order where a minority remains in power by violently repressing a majority. According to one explanation based on rational choice theory, the apartheid regime persisted because the oppressors were able to coordinate their actions while, at the same time, the oppressed were prevented from coordinating their opposition. If participating in a revolution when very few others participate is costly, and if it is not necessary that everyone participate in a revolt in order for it to succeed, then the oppressed will face a so-called free-rider problem. Since each individual member of the non-white population had an incentive to wait and see if enough others would stage a revolution, and since they were unable to coordinate their actions, nobody acted. We can follow Russell Hardin [1995, p. 29] and call this the *gunman theory of oppression*.[3]

According to this theory the oppressed South Africans were far from deceived. On the contrary, the oppressive social order persisted because the majority of the oppressed South Africans realised exactly what kind of situation they were caught up in and responded rationally to their situation.

With respect to the methodological commitments of the analytical Marxists there are a number of reasons to prefer the gunman theory to the ideology theory. First of all, the ideology theory of oppression involves functional ex-

[1]Cohen [2000, p. xxvi]. See also Cohen [2002] on bullshit.

[2]Literally, "Marxism without bullshit"

[3]We will discuss this theory at length in chapter 3.

planations. Among analytical Marxists much of the scientific failure of Marx's own theory was attributed to his use of function explanations. After all, among the natural sciences only biology employs functional explanations. And biology, unlike the social sciences, has access to an underlying well-confirmed law provided by Darwin's theory of natural selection.[1] Since functional explanations in the social sciences are controversial and since the gunman theory involves only well-accepted explanations in terms of purposive action, there are good reasons to prefer the gunman theory to the ideology theory of oppression. Another reason to prefer the gunman theory, is that at the few times the underlying mechanism of ideology is provided it often includes references to Hegelian Geists or Durkheimian organicism. The mechanisms offered by the gunman theory, on the other hand, refers only to familiar entities such as individual actions, beliefs, and desires. Since most of our other scientific theories do not involve societies as organisms or the existence of Hegelian Geists, and at the same time acknowledges the existence of individual actions, beliefs and desires, ontological parsimony gives us a reason to prefer the gunman theory over the ideology theory.[2]

These reasons are, however, not conclusive. Even if functional explanations and Geists are not acknowledged by our other scientific theories, the fact that some phenomena cannot be explained without functions can provide us with a reason to expand our list of acceptable forms of explanation. This can be compared to Emile Durkheim's [1979] argument for the existence of social entities in *Suicide*. He argued that since it was impossible to explain the change in suicide rates between 1856 and 1878 in terms of psychological facts, there must exist a social fact that caused the change. In other words, if our *best explanation* of the change in suicide rate uses social facts, then this gives us a good reason to believe that social facts exist. Similarly, if our best explanation of persistent oppression would be functional or involve Geists then we would have a good reason to accept functional explanation or believe that there are Geists.

Some proponents of the gunman theory do, however, claim that it is able to explain everything we need to explain about the persistence of oppression. [Rosen, 1996; Tännsjö, 2006] If they are right, then the fact that the ideology theory brings additional ontological baggage without adding explanatory power, provides us with a strong reason to favour the gunman theory over the ideology theory of oppression.

In a nutshell, the objection against the ideology theory of oppression can be formulated as follows:

[1]See, e.g., Elster [1982]. For some more general critique of functional explanations see Hempel [1994]. We will address these objections in chapter 5.

[2]See, e.g., Rosen [1996] and Cudd [2005].

1. It is possible to explain the persistence of all oppressive social orders without the ideology theory of oppression.

2. By accepting the ideology theory of oppression, we become committed to the existence of ontologically queer entities.

3. We should not accept a theory that commits us to the existence of ontologically queer entities unless this is necessary in order to explain something we are interested in explaining.

4. Therefore, we should not accept the ideology theory of oppression as an explanation of the persistence of oppressive social orders.

The aim of this book is to show that it is possible to accept both the ideology theory of oppression and the sound methodological principles that the analytical Marxists are committed to. We will accept premise 3, and reject both premises 1 and 2. We will argue that that there is nothing necessarily mysterious about the ideology theory of oppression. In fact, we will argue that the analytical Marxist's decision to jettison the ideology theory of oppression together with the "bullshit"-part of Marxism was to throw out the baby with the bathwater.

Before we turn to the task at hand, however, we need to say something about explanations. We will look at Bas van Fraassen's [1980] suggestion that explanations are answers to why-questions. It will be suggested that we have a good explanation of a phenomena if we have a *relevant* answer to the why-question. The problem is, of course, to specify what constitutes a relevant answer. This problem will not be solved. Instead, it will be argued that for our purposes it is enough that the ideology theory of oppression does not commit us to any queerer entities or explanations than the gunman theory. In section 1.2 the organisation of the book will be presented.

1.1 Explanations

To provide an explanation of a phenomenon, P, is to provide an answer to a why-question of the form "Why P?". To provide a good explanation is to provide a relevant and interesting answer to the same question. Well, so much for platitudes. In order to specify what a relevant and interesting answer consists of, let us begin with van Fraassen's [1980] rather minimalistic account of relevance and see if it can give us some criteria for good explanations. In the process we will see what the objections against the ideology theory of oppression are aimed at.

First, let us follow van Fraassen and call P the *topic* of a why-question. Natural scientists interested in global warning formulate theories in order to answer questions like "Why is the average temperature rising?", whereas social scientists might provide theories in order to answer "Why did the Copenhagen talks break down?" And in the case of theories of oppression, they can be said to be concerned with questions of the form:

1. Why does oppression of group X in society S persists?

Having identified the topic, however, is not enough for completely characterising a why-question. The problem is that when only the topic is specified a why-question allows for many different interpretations. Think of the following why-question "Why does the oppression of the citizens of North Korea by a small political elite persist?" It can be interpreted as asking why the North Koreans are oppressed and not the South Koreans, or why a small rather than a larger group is responsible for the oppression, or why the North Koreans are being oppressed and not treated to ice cream. All of the interpretations ask why is something (the topic, P_k) the case, rather than something else ($P_{i\neq k}$). In order to identify the correct interpretation of a why-question we can add a *contrast class*, C, of propositions that the topic is to be compared to. Thus, the purpose of offering an explanation is to answer the question "Why P_k is the case rather than $P_{i\neq k} \in C$".

It should be mentioned that one way of dismissing an explanation request, according to van Fraassen, is simply to deny the presupposition of the question "Why P_k rather than C?" This can simply be done by denying that P_k is the case. For example, the question "Why does the oppression of the workers persist (rather than ...)?" has the presupposition that the workers are oppressed. It is, therefore, possible to dismiss the explanation request just by denying that the workers are oppressed.

Returning to the task at hand of identifying why-questions, it seems as the pair $< P_k, C >$ is not enough to fully specify the question. According to van Fraassen, we need one last component to completely specify a why-question; that is to specify what kind of answers we are interested in. In other words, what kind of answers that are relevant to the question. van Fraassen does this by introducing a *relevance relation*, R, that has to hold between the topic-contrast class pair, and a proposition, A, in order for A to be part of an appropriate answer. If we manage to identify the relevance relation we can express the why-question with the help of the triple $Q =< P_k, C, R >$.

van Fraassen includes the relevance relation in order to specify the standards for an appropriate answer in a given context. Think of the question "Why did the match catch fire rather than fizzle?" and three possible answers: because...

1. A_1 = ... it was struck.

2. A_2 = ... the striking surface was made of sand, powdered glass, and a chemical called "red phosphorus." The head of a match was made of sulphur, glass powder, and an oxidising agent. When the match was struck on the striking surface of its box, the friction caused by the glass powder rubbing together produced enough heat to turn a very small amount of the red phosphorus into white phosphorus, which catches fire in air. This amount of heat was enough to start the chemical reaction that used the oxidising agent to produce oxygen gas. The heat and oxygen gas caused the sulphur to burst into flame which then caused the wood of the match to catch fire.[1]

3. A_3=... fish.

In the context of a child posing the question to a parent, then the relevance relation would pick A_1 as the appropriate answer. When posed to an engineering student at an exam, A_2 would probably be the most appropriate answer. Even A_3 could count as the appropriate answer if the question was posed as a surreal joke. Thus, we need the relevance relation in order to identify whether the why-question is part of, e.g., everyday discourse, scientific inquiry, or a surreal joke.

Let us focus on the cases where the question is posed as an explanation request and see what conditions the relevance relation impose on an explanation-answer in this context. Once the why-question has been identified as the triple $Q =< P_k, C, R >$, van Fraassen goes on to make some useful definitions. First, Q is said to have a *presupposition*, namely

(a) P_k is true,

(b) each $P_{i \neq k}$ is false, and

(c) there is at least one true proposition A that bears the relation R to the pair $< P_k, C >$.

The conditions (a) and (b) are said to be the central presuppositions of Q. Furthermore, the context in which the question is posed includes a body of background knowledge, K. Q is said to *arise* in a context if K entails the central presupposition, and does not entail the falsity of (c).

[1]This excellent answer was provided by the Science Theatre at Michigan State University, http://www.pa.msu.edu/sciencet/ask_st/092596.html, accessed on 13 May 2011.

In scientific contexts K is usually taken to consist of the state of science at the time the question is posed.[1] In other words, a statement that did count as explanatory in Newton's time might not count as explanatory today, and vice versa. Cohen [1978, p. 264] does, for example, point out that a presupposition of early modern physics included a principle forbidding action at a distance, and that Newton's laws of motion were not regarded as explanatory (even by Newton himself). Cohen goes on to point out that this principle was, by the mid-nineteenth century, abandoned.

As mentioned above, one way of rejecting an explanation request is to deny the presupposition of the question. For example, to deny that it is true that Swedish workers are oppressed (i.e., denying (a)). The central presupposition can also be rejected by denying that all other members of the contrast class are false, e.g., claiming that contrary to what is presupposed there exists a revolutionary movement in North Korea.

For our purposes, the most interesting way of rejecting the explanation request is to claim that K entails the falsity of (c). That is, to point out that given our present state of scientific knowledge there is no true proposition A that bears the relevant relation to the topic and contrast class. It can, for example, be claimed that the only answers to why the oppression of Swedish workers persists, involves the existence of a Hegelian Geist. Since our present body of scientific knowledge does not include such entities, there will be no appropriate answer to the question. Therefore, we should reject the explanation request. This is the kernel of Rosen's [1996] objection that will be discussed in chapter 6. A methodological individualistic version of this objection will be discussed in chapter 5.

If a question, Q, arise in a context, then a *direct answer* to it has the following form:

(*) P_k in contrast to the rest of C because A,

where the following conditions are met:[2]

(i) A is true, and

(ii) A bears R to $< P_k, C >$.

[1] See also Kitcher and Salmon [1987].

[2] van Fraassen [1980, p. 143] also adds the following two conditions:

(iii) P_k is true, and

(iv) no member of C other than P_k is true.

These are, however, fulfilled if the question arises. That is, if K entails (a) and (b), and if it does not entail the negation (c).

A is called the *core* of the answer, since the answer can be abbreviated "because *A*."

The problem with this account is, however, that it puts too few restrictions on the relevance relation to capture what we intuitively take to be a relevant answer to a why-question. Philip Kitcher and Wesley Salmon [1987, p. 319] show that if we do not put some additional restrictions on R, then for any true propositions A and P_k (and contrast class C where each member except P_k is false), there will be a why-question with P_k as its topic such that A is the core of the only direct answer to the question. Thus, they conclude, "if explanations are answers to why-questions, then it follows that, for any pair of true propositions, there is a context in which the first is the (core of the) only explanation of the second."

Kitcher and Salmon [1987, p. 322] illustrate the triviality result by offering an explanation of JFK's assassination in terms of astrology by setting R to a relation of astral influence so that only the answers that relate the position of the stars at a person's birth with her fate become relevant. Properly constructed, the proposition describing the position of the stars at JFK's birth will then become the only core to the question of why JFK was assassinated on November 22 rather than on November 21 or November 23. The problem is not that van Fraassen's account allows this to count as a good explanation, but rather that it allows it to count as an explanation at all. Given the state of 20th-century science, the appropriate response to a why-question where R specifies astral relations is that it should be rejected.

The moral Kitcher and Salmon draw from these examples is that, in order to escape an "anything goes"-conclusion about scientific explanation, more restrictions on R are needed. The question of whether an answer to a why-questions counts as a good scientific explanation, would thus amount to whether it bears the *relevant* relevance relation to the question's topic.

What counts as the relevant relevance relation does, however, bring us to the old question of what counts as a scientific explanation. One way of modifying van Fraassen's account is to supplement it with Hempel's [1965, p. 247] deductive-nomological (DN) model of scientific explanation. Although there are many problems with the DN-model, we will begin our investigation by assuming that any explanation that conforms to it is a good explanation.[1] According to the DN-model an explanation consists of two sets of propositions: an *explanandum*, describing the phenomenon that is to be explained, and an *explanans*, the set of propositions describing that which is to explain. The explanans can further be divided into a set of propositions describing natural laws, and a set of proposition describing initial conditions. The explanans is

[1] We will discuss some of these problems in chapter 5.

then said to *explain* the explanandum if all propositions in the explanans are true, and the explanandum can be logically deduced from the explanans. For example, if we want to explain why a match caught fire with the help of the DN-model we would need two premises and a conclusion:

1. Whenever a match is struck, it catches fire,

2. the match is struck,

3. therefore, the match catches fire.

Although the first premise is not a natural law, it can probably be shown to follow from some natural laws and initial conditions. The second premise describes the relevant initial conditions. Together they constitute the explanans. The third proposition describes the explanandum. Since the explanandum can be deduced from the explanans (and if we assume that that they are true), then we have an explanation of why the match caught fire.

In van Fraassen's model of explanation this would amount to a relevance relation, R, that picked out all As as direct answers that allow the derivation of P_k from A (and perhaps some additional premises from K). Under the Hempelian reading $< P_k, X >$ is the explanandum whereas A is the explanans. Since there is no natural law of astral influence, the attempted explanation of Kennedy's murder, in terms of the position of the stars at his birth, is blocked.

We will have occasion to modify the account of a good explanation as we go along. We will have to do this not only when we investigate the ideology theory of oppression, but also when we try to spell out the gunman theory of oppression in chapter 3. Many of the propositions included in the explanans of the gunman theory of oppression are obviously false, yet they seem to offer good explanations. In chapter 5, we will investigate whether the mechanism-account of explanation, motivated by the shortcomings of the DN-model, is compatible with the ideology theory of oppression.

The general strategy will be to use the form of explanation used by the gunman theory as a benchmark when we evaluate the ideology theory of oppression. All that we will require of the ideology theory is that it does not presuppose any stranger relevance relations, or does not commit us to any queerer entities than the gunman theory.

1.2 Organisation of the book

The book will be organised as follows. In chapter 2 we will introduce the explanandum that the theories of persistent oppression should be able to explain. We will begin by discussing oppression in general and argue that when certain

criteria are violated social orders tend to be oppressive. We will then proceed to claim that people (as an empirical matter of fact) tend to dislike living in social orders where these criteria are violated. The chapter will conclude with a description of three real-life persistent oppressive social orders: tyranny in North Korea, economic inequality in USA, and gender inequality in Sweden.

Chapters 3 will begin by introducing the basics of rational choice and game theory. We will then proceed to discuss the gunman theory of oppression at length and show how it can provide an elegant mathematical model that can be used to explain the persistent oppression in North Korea (and other tyrannies).

In chapter 4, we will show that there are some interesting aspects of oppression that the gunman theory of oppression fails to explain. We will provide a modest version of Durkheim's argument and show that rational choice models fail to explain persistent economic inequalities in USA and persistent gender inequalities in Sweden. We will suggest that ideological beliefs and norms can be used to offer a straightforward explanation of these cases. We will also provide a more formal statement of the ideology theory of oppression where the functional component will be made explicit. The chapter will conclude with a discussion of some preliminary objections against this theory.

If the ideology theory of oppression has a functional component, and if functional explanations are as controversial in the social sciences as it has sometimes been claimed, then we might still have strong reasons to steer clear from the ideology theory of oppression. The benefit of explaining the new cases of persistent oppression may not be enough to compensate for the cost of having to include the new ontological entities. Many have worried that the use of functional explanations in the social sciences force us to accept ontologically queer entities. In chapter 5, we will show that this worry has been greatly exaggerated. More specifically, we will defend Cohen's [1978] account of functional explanations against the so-called *no-mechanism* objection. According to this objection, we need epistemic access to an underlying causal mechanism in order to be justified in believing a functional explanation. We will show that this argument presupposes that all social-level causes should be reduced to individual-level causes. We will argue that this requirement is too strong. A more reasonable demand is that all social-level causes used in an explanation should *supervene* on individual-level causes. It will be shown that if supervenience is accepted, then functional explanations will be acceptable as long as they give us interesting answers to the why-questions we are interested in. Finally, we will show that, in some cases, there are reasons to believe that the individual-level explanations cannot offer as interesting explanations as functional explanations.

However, showing that functional explanations are acceptable in the social sciences does not show that functional explanations of persistent oppression

are acceptable. It is more or less obvious how some functional explanations in the social sciences can be elaborated in terms of subvenient micro-causes. For example, a functional explanation of why car factories tend to operate on a large scale is acceptable, since it not only gives us additional interesting insights into the car industry, but also because it is relatively easy to identify the possible subvenient causes. The explanation might be elaborated, for example, in terms of the purposive actions of the car factories' managers, or in terms of some evolutionary mechanism that eliminates unprofitable small-scale factories. However, the problem with the ideology theory of oppression might be that there are *no conceivable* subvenient causes that can realise the suggested function. Michael Rosen [1996], who accepts the use of functional explanations in the social sciences in general, has argued that the problem with the ideology theory of oppression is that there is no way to reconcile it with our standard ontological commitments. According to Rosen, all elaborations of the ideology theory of oppression contain references to, e.g., Hegelian Geists or society as an organism. In chapter 6, we will show that this is wrong and we will offer some demystifying elaborations of the ideology theory of oppression.

Although the task of offering a full explanation of why oppression persists will be left for the social scientists, three possible elaborations of the ideology theory will be offered in chapter 6 in order to indicate how it can be demystified. The first two are based on the suggestion by Cohen [1978] that there are traces of a Darwinian mechanism in Marx's writings. The third is a synthesis of Cristina Bicchieri and Yoshitaka Fukui's [1999] information cascade model with trendsetters and the famous propaganda model of Edward Herman and Noam Chomsky [1988]. Although all three models have their merits, we will argue that the last is best suited for the analytical Marxists who want to use the ideology theory of oppression while retaining their commitment to rational choice theory.

Finally, chapter 7 will conclude the book by attempting to offer an answer to Vladimir Lenin's [1988] question: "What is to be done?" We will begin by suggesting a statistical method for discovering whether an oppressive social order is upheld by ideological beliefs or norms. We will then argue that if oppressive social orders are in fact upheld by ideology, then our chances of getting rid of the oppression through democratic means might be slim. After all, if the capitalist system produces the belief that, e.g., tax cuts are preferable to increased taxes and free healthcare, then it will be futile to try to get elected with that political agenda of introducing free healthcare. Thus, Lenin's suggestion that we need a revolutionary vanguard to lead the struggle may be the only solution. If the ideology theory of oppression is correct, then the most direct way of getting rid of the ideological beliefs will be to destroy, what Marx

called, the economic base. We will, however, end by pointing out that whether we should go this far is, in part, an empirical question and, in part, a question for moral and political philosophy. After all, if the costs of revolutionary action exceeds the benefits of getting rid of the oppression, then we might be forced to draw the pessimistic conclusion that maintaining the oppressive status quo is preferable to a violent revolution.

2. Oppressive social orders

2.1 Introduction

In an oft-quoted passage William Reich formulates a problem for social psychology as follows:

> What has to be explained is not the fact that the man who is hungry steals or the fact that the man who is exploited strikes, but why the majority of those who are hungry *don't* steal and why the majority of those who are exploited *don't* strike. [Reich, 1970, p. 14]

According to Reich, any situation where a group of people are being exploited without going on strike is prima facie mysterious and demands an explanation. The same goes for a situation where a group of people are being deprived of food and do not steal in order to satisfy their hunger. It is mysterious because people tend to strongly dislike going hungry and being exploited. In the light of Terry Eagleton's [2007, p. 27] remark that "the majority of people have a sharp eye to their own rights and interests, and most people feel uncomfortable at the thought of belonging to a seriously unjust form of life," we can generalise Reich's claim to include all forms of injustices and oppression. So, to paraphrase Reich, what needs to be explained is not the fact that the oppressed man revolts, but rather why the majority of those who are oppressed *don't* revolt.

However, to evaluate our theories of persisting oppression we will need some concrete examples of oppressive social orders. In order to identify these social orders, it will first be necessary to spell out the mystery of oppression in some more detail.

The purpose of this chapter is to formulate the mystery of oppression and then to provide some examples of real social orders that fit this description. In order to do so we will first have to say something about oppression in general. We will notice that most philosophers who discuss oppression define it in terms of injustice. However, since people often disagree about normative terms, like injustice, we will instead suggest a set of criteria that can be used to pick out the social orders people *tend to consider* unjust and oppressive. Once we have spelled out the mystery of oppression, we will have a look at three oppressive social orders that fit the description.

In section 2.2, we will spell out the mystery of oppression and answer some immediate objections. In section 2.3, we will have a look at some of the definitions of oppression from the philosophical literature. Most definitions are given in normative terms claiming that oppression implies injustice. We will point to some possible problems for our project of defining oppression in normative terms and then suggest that we can avoid these problems by focusing on the social orders that people under normal circumstances tend to consider unjust. In section 2.4, we will spell out the criteria that allows us to pick out these social orders. Once we have formulated the criteria we will, in section 2.5, move on to introduce the real-life examples of oppressive social order that our theories of persistent oppression should explain. Finally, section 2.6 will conclude the chapter.

2.2 The mystery of oppression

What calls for an explanation is that within some social orders there are groups that are oppressed, this oppression is bad for the members of these groups, and they have (in some sense) the ability to exchange the oppressive social order for a less oppressive social order. These claims can be summarised as follows:

> THE MYSTERY OF OPPRESSION: There exists a persistent social order, O_1, a possible persistent social order, $O_2 \neq O_1$, and a group X, such that
>
> 1. X is more oppressed in O_1 than in O_2,
> 2. O_2 is better than O_1 (in terms of welfare) for the members of X, and
> 3. X has the ability to exchange O_1 for O_2.

Before proceeding to the meaning of "oppression," let us take care of some preliminary objections against this formulation of the mystery of oppression. For example, it might seem as persistent social orders where either 1 or 2 are satisfied in conjunction with 3 are mysterious. After all, if people would be better off under an alternative social order, and can do something about it, why do they not do so? Similarly, if people are oppressed under one social order, and can do something about it, why do they not do so?

To answer the first question, it is sufficient to consider that people often accept distributions of welfare even if there exist alternative distributions where they could have received more welfare. For example, many (although not all) rich people in Sweden and other welfare states accept that a significant share of their wealth and income is redistributed to the poor. They may do this because

they believe that the redistribution is just, or because they accept the democratic process that have lead up to the decision to redistribute. A social order may, in other words, have certain redeeming qualities that make people accept it although there exists an alternative social order where they would be better off. In order to rule out the existence of such redeeming qualities, we add the claim that the members of X are more oppressed in O_1 than in O_2.

Similarly, to answer the second question we may note that people at times accept oppressive social orders if the less oppressive social order seriously lowers their welfare. After all, people cannot eat their freedom of speech and thus may be willing to put up with some oppression to receive their daily bread. In order to rule out this immediate solution to the mystery of oppression, we add the claim that the less oppressive social order, O_2, is better in terms of welfare for the members of X than O_1.

It might, of course, be the case that some social orders where either 1 or 2 are satisfied in conjunction with 3 are mysterious. The oppression may, for example, be so grave that a potential decrease in welfare does not seem to be enough to explain the lack of civil unrest. Or the potential increase in welfare may be so high that the fact that the members of X are not oppressed does not seem to explain their inactivity. We will, however, focus on the really mysterious cases where both 1 and 2 are satisfied. If the theories of persistent oppression can offer explanations of the really mysterious cases, then they can probably explain the less mysterious cases as well.

It might, however, be objected that conditions 1 and 2 are biased in favour of the ideology theory of oppression. After all, the gunman theory relies on rational choice theory, and therefore it is not interested in whether the members of X *are* oppressed or whether the less oppressive social order *would be better* for them, but rather whether the members of group X *believe* that they are oppressed and whether they *prefer* the less oppressive social order.

This will, however, not be a big problem, it only means that the gunman theory of oppression is able to explain more phenomena than we are interested in. For our purposes, the gunman theory of oppression will be relevant if we assume that there is a connection, on the one hand, between a social order being oppressive for X, and the members of X believing that the social order is oppressive; and, on the other hand, a less oppressive social order being in the interest of the members of X, and the members of X preferring the less oppressive social order.

However, if we would formulate the mystery of oppression in terms of beliefs and preferences, our theory would become biased in favour of the gunman theory of oppression. After all, the proponents of the ideology theory of oppression often try to resolve the mystery by explaining why people who *are* oppressed do not *believe* that they are oppressed, or why they do not *prefer* a

social order that clearly is in their *interest*.

Finally, as condition 3 is stated, it might be objected that it is strange to attribute an ability to a group. This claim should, however, be understood in the same way as the claim that a group of friends has the ability to push a car over a hill. The claim that a group of friends has the ability to push a car over the hill does not imply that any individual member of this group has this ability, nor does it imply that the group has any ability over and above the individual members. It only means that if they would work together and exercise their individual abilities and push at the same time, they would successfully push the car over the hill. Similarly, the claim that a group of oppressed has the ability to exchange a social order for another does not imply that any individual member of this group has the ability to overthrow the oppressive social order, nor does it imply that this group has some mystical power over and above the abilities of its members. It only means that if the individual citizens would exercise their individual abilities together, then they would successfully exchange the present social order for another. For example, claiming that the North Korean people have the ability to overthrow Kim Jong-Un does not imply that any individual North Korean has this ability or that the group of North Koreans has some mysterious power. If the North Korean citizens exercised their ability to protest and revolt at roughly the same time, then they would successfully overthrow Kim Jong-Un's regime. As we will see in chapter 3, the gunman theory of oppression usually tries to explain the mystery of oppression by pointing to the failure of the oppressed to coordinate their actions.

2.3 Oppression

It is a bit awkward to discuss theories of persistent oppression without having a clear understanding of what we mean by oppression. There is probably some overlapping consensus about the social orders that are oppressive. Almost everyone would agree that, for example, the North Korean people are oppressed, African Americans were oppressed under the Jim Crow legislation, and women are oppressed under the Taliban rule in Afghanistan. There are, however, other social orders where the oppression of some groups is disputed. For example, some, but not everyone, agrees with the claims that African Americans continue to be oppressed also after the abolishment of the Jim Crow laws, women are oppressed in Sweden, or workers are oppressed in capitalist societies.

It is possible to find one explanation as to why there is so much disagreement about which situations should count as oppressive if we look at the attempts to define oppression. Beginning with the Oxford English Dictionary oppression is

> prolonged cruel or unjust treatment or exercise of authority, control, or power; tyranny; exploitation.

The most conspicuous part of the explication is probably that oppression might involve *unjust* treatment. A similar normative component can be found in most attempts to define oppression in the philosophical literature. Ann Cudd [2006, p. 25] does, for example, argue that the use of *unjustified* coercion is a necessary condition for oppression. Sally Haslanger [2004, p. 98] states that x oppresses y only if y suffers *unjustly or wrongfully* under x. And Alan Wertheimer [1996, p. 18] claims that x oppresses y when x deprives y of freedoms or opportunities to which x is *entitled*.

Even if most philosophers agree that oppression involves unjust treatment, there can still be disagreement about the criteria that must be fulfilled in order for some treatment to count as unjust. For example, a Marxist and a libertarian can agree that all oppression is unjust, but disagree about whether workers are oppressed in a capitalist society since the former but not the latter considers the treatment of the workers as unjust. The normative component of oppression, thus, accounts for some of the disagreement.

Let us return to this problem once we have had a look at a more serious objection that can be raised against the attempt to define oppression in normative terms. If the theories of persistent oppression claim that there are social orders in which a group is oppressed, then the theory implies that there are true normative claims and as such it is incompatible with certain meta-ethical theories. Consider, for example, error theory that implies that all moral claims are false. If error theory is correct, then the claim that Palestinians are oppressed by Israeli settlers on the West Bank is false, therefore what is mysterious about the persistence of this (and all other oppressive) social orders would disappear.

It would indeed be surprising, if our specification of the mystery of oppression implies that the mystery would disappear if we discover that error theory is the correct meta-ethical theory. After all, error theory does not seem to have anything to do with the mystery of why (and how) a small group of settlers can dictate the conditions for a large group of Palestinians on Palestinian territory.

We can get around these problems by focusing on the criteria that are violated by the social orders people (under normal circumstances) consider to be unjust. It could after all be argued that what is mysterious about oppressive social orders is not so much that they are oppressive, but rather that these social orders violate certain criteria that people tend to cherish.

Thus, an error theorist can think of an oppressive social order as a social order that violates certain well-accepted criteria of justice. What makes the persistence of this social order mysterious is simply that people tend to strongly dislike living in societies that they consider to be unjust. The disagreement between the libertarian and the Marxist is "resolved" in a similar fashion. After

all, it does not really matter who has offered the correct account of injustice or oppression. What matters is whether people tend to dislike living in societies that violate some such criteria. Since it is an empirical question whether people tend to dislike these social orders, it is possible to formulate the problem of persistent oppression without having to take a stand on normative or meta-ethical issues.

We will, in section 2.4, suggest some criteria that people tend to demand that their social orders should satisfy.[1] We will see that it is possible to treat the claim that a social order is oppressive as having no normative or meta-ethical implications. Under this interpretation, an oppressive social order is a social order that lacks certain properties that people (under normal circumstances) tend to demand.[2]

Having offered these caveats, let us continue by making some remarks about some other features of oppression. First, most definitions include a condition that the treatment has to be systematic or institutionalised. Cudd's [2006] definition does, for example, require the presence of an institutional practice defined as "formal and informal constraints such as law, convention, norms, practices, and the like." [Cudd, 2006, p. 20] It is necessary to include a similar condition in order to be able to distinguish between isolated injustices and oppression. Although it does not really matter which word we use, we will opt for the use of the word 'systematic' instead of 'institutional' since it allows us to speak about oppression where it is ordinarily inappropriate to speak about institutions.

Second, unjust and systematic treatment cannot be the whole story of oppression. If it were, then we would perhaps have to say that Robin Hood was oppressing the rich when he systematically robbed (that is if stealing from someone is to treat them unjustly) the rich to give to the poor. We can avoid this implication by demanding that the oppressed do not stand in a relation of dominance to the oppressors in order for the treatment to count as oppression.

Third, oppression might need to be contextualised so that one group, x, can both oppress and be oppressed by another group, y, in the same society. We can, for example, think of a social order where men oppress women on the labour market, whereas women oppress men at home. So oppression might need to be defined as being relative a certain context.

Let us sum up what we have said and offer the following definition of an

[1] We will suggest two criteria that can be derived from John Rawls' [1999] second principle of justice that includes a) the difference principle and the b) the principle of fair equality of opportunity.

[2] By saying that a person would make these demands under normal circumstances we try to exclude situations where people are subject to thought control, coercion, or severe bias.

oppressive relation:

> Oppression: x is OPPRESSED by y in context c if, and only if, y systematically treats x unjustly and x does not stand in relation of dominance to y in c.

That the oppressive social orders we will investigate will imply systematic treatment will not be an issue. Since we are interested in *persistent* social orders, whatever treatment group X is subjected to will more or less by definition be systematic.

Furthermore, a relation of dominance is a somewhat technical term, and a lot more could be said about it. However, in the social orders we will investigate it should be sufficiently clear that the group that is treated unjustly does not stand in a relation of dominance to their putative oppressors. Similarly in the social orders we will investigate, there does not seem to be any contexts in which the oppressed become oppressors and therefore (in some sense) become compensated.

What will be at stake, however, is whether the social order implies that some group is treated unjustly. Or, to be more precise, whether the social order violates criteria of justice that people tend to accept. After all, as we mentioned in section 2.2, if a social order is not (normally considered to be) unjust or oppressive, then it might not be much of a mystery that the order persists.

2.4 Unjust treatment

As we mentioned in section 2.3, the claim that X is oppressed seems to be normative in the sense that it implies that X is treated unjustly. We also pointed out that this made the truth of premise 1 (X is more oppressed in O_1 than in O_2) incompatible with certain meta-ethical and normative theories. We did, however, also argue that we could resolve this problem by focusing on criteria of justice that people tend to accept. This will allow us to retain what is mysterious about persistent oppressive social orders even if it turns out that what we had the wrong idea about justice, or if we discover that there are no true claims about justice at all. As long as people tend to dislike social orders that lack certain properties, the mystery of oppression will remain intact.

There are many theories of justice that can provide us with candidate criteria. We could, for example, opt for Robert Nozick's [1974] libertarian theory of justice and focus on entitlement. This would give us a criterion that forbids distributions that are the result of unfree exchanges, no matter how efficient or equal these distributions would be. We could get an altogether different criterion if we opted for Marx and Engel's famous slogan: "from each according to his ability, to each according to his need." [Marx and Engels, 2001, p. 20]

This would result in a criterion that forbids all unequal distributions, no matter how they came about or how efficient they would be. A criterion based on the Marxist slogan would probably rule out most distributions as unjust that the libertarian criterion would count as just, and vice versa.

For our purposes, however, both of these conditions would be problematic since it is far from obvious that people tend to dislike social orders where they are violated. Most societies do, for example, have laws against certain free exchanges. Exchanges involving sexual services, drugs, and euthanasia are not only illegal, but also frowned upon in many societies. Furthermore, most societies allow people to keep most of what they earn even if the resources would satisfy someone else's needs better, and there seems to be good reason to believe that people would strongly oppose a complete redistribution based on need. It is, of course, possible that those who oppose policies based on these theories of justice are victims of ideology in the sense we will discuss in chapter 4. However, in order to avoid handicapping the ideology theory of oppression we should try to find some theory that holds some more immediate intuitive appeal.

Let us, therefore, consider one of the most influential theories of justice of the 20th century: John Rawls' [1999] theory from *A Theory of Justice*. In a sense Rawls' theory of justice combines the intuitive appeal of liberalism, socialism, and consequentialism. The second principle of justice states that

> social and economic inequalities are to be arranged so that they are both
>
> (a) to the greatest expected benefit of the least advantaged, and
>
> (b) attached to positions and offices open to all under conditions of fair equality of opportunity.

The first condition, the so-called *difference principle*, appeals to similar intuition as the Marxist slogan. That is, we should take the needs of the worst off into account when we distribute resources. It does, however, also take into account that a full redistribution would have bad consequences since it creates an incentive to free-ride on the efforts of others. The second condition, sometimes called the *fair equality of opportunity principle* (FEO), appeals to similar intuitions as liberalism: everyone should (in some sense) have the same chances of succeeding.[1]

There are many reasons to doubt that the second principle of justice successfully captures all our intuitions concerning justice.[2] It does, however, seem

[1] More about this below.

[2] See, e.g., Robert Nozick's Wilt Chamberlain objection, and G.A. Cohen's [2008, ch. 4] attempt to rescue justice from the difference principle.

to capture some of the more common intuitions in the sense that it describes the conditions under which people are willing to accept unequal social orders.

Before we discuss this, let us attempt to extract two more concrete criteria of justice from Rawls' principle. We will begin by extracting a criterion from the second condition (fair equality of opportunity) and then move on to discuss the difference principle at some length. When discussing the possibility of extracting a criterion from the difference principle, we will get the opportunity to say something about the relationship between welfare, interests, and motivation. Having some understanding of the relation between what makes a person's life go well and what motivates her will be important when we discuss the ideology theory of oppression in chapter 4.

2.4.1 Equality of opportunity

If we want to extract a criterion that people tend to accept from condition (b), it might be tempting to spell it out in terms of formal equality of opportunity: everyone should have the same legal rights to access all privileged social positions within a social order. On this reading the condition will be satisfied if there are no formal laws that prevents the members of some group to hold certain privileged positions in a society. One example of a social order where the formal version of the condition was obviously violated is South Africa under apartheid, where non-whites were prevented from attending regular schools, voting, and holding office. A present-day social order that violates this condition is Saudi Arabia where women, among other things, are prevented from voting and holding office.[1] The formal condition gets some intuitive support from the fact that most of us would judge these social orders to be oppressive.

However, if the condition demands only formal equality of opportunity, then it does not seem to rule out many social orders we normally consider to be oppressive and unjust. The formal equality of opportunity condition does, for example, not rule out the American South as unjust under the Jim Crow legislation that was supposed to keep white and black Americans "separate but equal." Since there was no law that prohibited African Americans (qua African Americans) from voting or holding office the Jim Crow legislation did not violate the formal version of the equality of opportunity condition. What the laws did, however, was to prevent everyone who failed a literacy test from voting (and thus holding office),[2] while at the same time demanding schools to be segregated. Since schools for African Americans tended to be

[1] See for example Asmaa Al-Mohamad [2008].

[2] Arizona and Washington demanded that electors pass a kind of literacy test, whereas Maine, Oklahoma, Oregon, and Wyoming required electors to be able to read the state Constitution.

poorer than schools for whites, the African American population tended to score lower on the literacy tests than the white population. Consequently, the African American population was to a larger degree than the white population prevented from voting and holding office.

In order to capture the intuition that obviously discriminating laws (such as the Jim Crow laws) are unjust, Rawls suggests that "fair" should require more than only formal equality of opportunity.

> Offhand it is not clear what is meant [by fair], but we might say that those with similar abilities and skills should have similar life chances. More specifically, assuming that there is a distribution of natural assets, those who are at the same level of talent and ability, and have the same willingness to use them, should have the same prospects of success regardless of their initial place in the social system. [Rawls, 1999, p. 63]

On this interpretation of fair equality of opportunity, a social order has to provide individuals at the same level of talents and abilities the same prospects of success. Since the literacy tests and segregated schools effectively reduced the African Americans' chances of succeeding, the states that applied the Jim Crow laws clearly violated this condition.

However, the condition seems to demand more from social orders than to simply refrain from passing and upholding discriminating laws. It does, for example, seem to demand that social orders provide opportunities to members of disadvantaged groups so that they get the same prospects of success as members of advantaged groups.[1] In addition to avoiding passing and upholding laws that are biased against an already disadvantaged group, this may be done by providing free or subsidised education and healthcare for the worst off in a society.

Although there is something intuitively appealing in demanding that social orders should compensate the disadvantaged, it is by no means uncontroversial. Although most would agree that social order should not pass laws that are biased against already disadvantaged groups, not everyone agrees that the government should actively compensate for existing inequalities. Furthermore, it might be argued that the difference principle (which we will look at shortly) captures some of the underlying intuitions that the least well off should be compensated.

We will, therefore, assume that the fair equality of opportunity condition demands something in between formal equality and equal prospects of suc-

[1] The condition may also be satisfied by preventing the members of the advantaged group from using their privileges. This would be a sort of levelling-down solution to the problem.

ceeding. The criterion will demand that in order for the fair equality of opportunity condition to hold, in addition to not passing laws that explicitly discriminate against the members of a group, social orders must avoid passing and upholding laws that are biased against already disadvantaged groups. This criterion will allow us to classify social orders as unjust if they demand that citizens pass literacy tests, have certain levels of income, or own land in order to vote and hold office. Whether it would count a social order as unjust if it demands unsubsidised tuition fees for university studies is unclear. For our purposes it will not be necessary to take sides on this issue.

Let us say that the fair equality of opportunity criterion is satisfied by a social order *A* if:

> FAIR EQUALITY OF OPPORTUNITY: the social and economic inequalities in *A* are arranged so that they are attached to positions and offices that nobody is prevented by law (directly or indirectly through a systemic bias) from attaining.

Although the formulation of the criterion is far from perfect, it will suffice. What is important for our purposes is that most people tend to dislike living in social orders where the condition is violated.

2.4.2 Maximin and welfare

Let us now turn to the difference principle that demands that economic inequalities are arranged so that they are to the greatest expected benefit of the least advantaged. For Rawls, the greatest expected benefit is supposed to be measured in *primary goods*, goods that one needs to realise any life plan, no matter what that plan might be. Primary goods are the goods we need in order to, e.g., run for mayor, studying Classics at Oxford, becoming a doctor, or becoming filthy rich. Some examples of, what Rawls considers, primary goods are basic rights and liberties, income and wealth, and recognition by the social institutions. Although we do not have to adopt his list, it provides us with at least one possible specification of expected benefits that lack any meta-ethical implication whatsoever: income and wealth. It is, furthermore, possible that rights, liberties, and recognition can be interpreted so that they do not carry any meta-ethical implications either.

For now, however, let us just note that Rawls' list of primary goods provide us with one example of a so-called *objective list*-specification of *welfare*. In what follows, we will use the terms welfare and wellbeing interchangeably. We will also talk about a person's interests and say that something is in the interest of a person if the satisfaction of this interest would increase the person's welfare.

Welfare is usually taken to be a measure of how good a person's life is for that person.[1] Thus, the welfare of a person's life can be contrasted to its aesthetic or ethical value. On the one hand, there are very dramatic lives, such as Hamlet's, that have high aesthetic values while at the same time being bad for the person whose life it is. On the other hand, there are very dull lives that lack aesthetic value altogether, but seem to be very good for the person who lives it. Similarly, the slogan "nice guys finish last" teaches us that a life with high ethical value can regrettably be (and perhaps often is) bad for the person living it.

By focusing on welfare, we are provided with an immediate connection between oppression and con-attitudes. If being oppressed detracts from a person's welfare, and if people tend to care about their welfare,[2] then it is indeed a mystery that people put up with being oppressed.

However, that people tend to care about their own welfare does not mean that people are always motivated to maximise their own welfare. After all, people are at times motivated by altruistic and moral considerations. Neither does it mean that people succeed in maximising their own welfare even when they try to. Apart from the obvious reasons, that they fail to realise their plan, people can be mistaken about what makes their life go well and therefore form the wrong plan. For example, Fidel might believe that taking the bus will be the quickest way to get to work, when, as it happens, an accident has blocked the road. Contrary to what Fidel believes, taking the bus will cause much frustration and lower his welfare.

People may also be systematically mistaken. For example, many people may smoke because they believe that this is good for them (in terms of pleasure, sense of freedom, or whatnot), when smoking is in fact bad for their health and therefore lowers their welfare. However, since people tend to care about their own welfare, some additional explanation will be needed to claim that people are systematically mistaken about what is in their interests. In the case of smoking an explanation could consist of claims about advertisement on behalf of big tobacco, or claims about the addictiveness of nicotine and psychological mechanisms causing people to rationalise their addictions, or of some combination thereof. The challenge for the ideology theory of oppression, as we shall see, is in part to explain why people are systematically mistaken about what makes their lives go well.

Before having a look at some theories of welfare, let us formulate the difference principle as a so-called maximin criterion in terms of welfare. The maximin criterion is satisfied by a social order *A* if:

[1]For a more extensive discussion on welfare, see Wayne Sumner [1996].

[2]Social indicator studies (and common sense) have shown that people do indeed tend to care about their own welfare. See, e.g., Sumner [1996, p. 150ff].

> THE MAXIMIN CRITERION: there is no other feasible social order B, where the worst off individual in B has a higher level of expected welfare than the worst off individual in A.

That is, the maximin criterion will be violated if there exists an alternative social order where the worst off are better off in terms of welfare than they are in the present social order.

For example, assume that the two groups X and Y live in social order A and that X has a higher level of welfare than Y. Furthermore, assume that X would be better off (in terms of welfare) and Y much worse off in the alternative social order B. If the members of X accept the maximin criterion, then they would refrain from opting for social change. Many relatively rich people who support high taxes and foreign aid are probably motivated by similar considerations. That the present social order satisfies the maximin criterion can be said to be a redeeming quality of the social order in the eyes of the relatively rich members of X.

Although the maximin condition is widely cited it should be pointed out that it can be objected that it is too strong. After all, in order to satisfy the condition a social order would have to give priority to its worst off members. Therefore, if the worst off members of a society need a lot of resources to get a very small increase in welfare, then the maximin condition will require a society to redistribute resources until the worst off cannot get any more welfare, or until everyone are on the same (low) level of welfare. In healthcare ethics, this is sometimes referred to as the "bottomless pit argument."

In order to save the maximin criterion from the bottomless pit argument, we could qualify the criterion by including some appeal to total welfare. We could, for example, add a clause that exempts a social order from increasing the welfare of the worst off if it would be extremely costly in terms of total welfare. We could also avoid the problem by prohibiting redistributions that lead to a levelling down of welfare.

For our purposes, however, it will be enough to note that in the cases of putative oppression we will describe in section 2.5, none of the alternative social orders will involve a decrease of total welfare or a levelling down of individual welfare. When we do not have to worry about bottomless pits and similar problems, then the maximin criterion seems to capture our intuitions about giving priority to the worst off.

Before we turn to the cases of persistent social orders, let us say something about welfare. Although it is tempting to offer an explication of welfare, the best thing we can do is probably to leave it unanalysed. For our purposes it does not really matter which theory of welfare is correct, as long as they allow for the possibility that people can be mistaken about what is in their interests. Let us, therefore, indicate how the three major theories of welfare

can account for such mistakes. The three major theories of welfare are the *hedonistic*, *desire satisfaction*, and the *objective list* account of welfare.[1]

According to the hedonistic account, a person's welfare is a matter of her pleasurable and painful experiences. This includes, but is not restricted, to bodily pleasures and pains. Hedonistic welfare can also include intellectual pleasures and pains, such as solving a complex mathematical problem and the frustration of not being able to finish one's dissertation in time. On this account, someone would be mistaken if she believed that something that did not maximise her pleasure (properly understood) was in her interest. For example, a person who believes that it is in her interest to dedicate her life to muzak and potatoes will be mistaken if muzak and potatoes do not give her the highest possible pleasure.

The desire satisfaction account treats welfare as a function of the desires, wants, or preferences being satisfied or frustrated. The more desires a person satisfies, the higher is her welfare. Proponents of the desire satisfaction theory usually impose further restrictions on the desires that when satisfied increase a person's welfare. It has, for example, been suggested that in order to be welfare-increasing a desire should be well-informed, autonomously formed, about one's whole life, or held under some ideal conditions. Once these restrictions have been imposed it is easy to account for mistakes about welfare. On the well-informed account, a person would be mistaken if she believed that her welfare was maximised when she acted on a desire that proved to be ill-informed. For example, a person's decision to begin to smoke in order to satisfy a desire of belonging to a group of smokers seems to be an example of such a mistake. Had she been well-informed, the desire to belong would have been silenced by the well-informed desire for being in good health.

Finally, the objective list accounts of welfare are made up of lists of things that are good or bad for a person irrespective of her attitudes towards them. We have already been acquainted with one example of an objective list in Rawls' primary goods. On Rawls' account, a person will have a high level of welfare if she possess certain basic rights, has a high income, and is recognised by the social institutions. On other lists, we find other good things, for example, the development of one's abilities, knowledge, appreciation of beauty, good health, nourishment, personal security, etc. Making mistakes on the objective list account is relatively straightforward, it is enough that a person believes that something that is not on the list is good for her. A person would be mistaken if she, for example, believed that hedonic pleasure is welfare-increasing when it is not on the list that turned out to be true.

We do not, however, need to go into the details of the theories of welfare.

[1]The names are also used by Derek Parfit [1987] to categorise theories of self-interest.

We do not, for example, have to discuss which analysis of welfare is the most plausible, or how their proponents have defended and motivated them. For our purposes it is enough to note that most of us have some intuitions about welfare. That is, we have some relatively clear ideas about when a person's life is good for that person.

Furthermore, there seems to be some overlapping consensus about what makes a life good for a person. It is very likely that most of us would, for example, agree that all other things being equal a person's life is better for her if she is in good health, has a high income, enjoys freedom of speech, is free to worship, is free to choose her partner, is in pleasure rather than pain, has her well-informed desires satisfied, etc. We will, therefore, assume that a social order that does not distribute the items on this list so that it satisfies the maximin criterion will be unjust, and therefore (if it satisfies the other conditions) oppressive.

If someone objects to this assumption and claims, for example, that it is false that a person's life tends to go better when she can freely choose her partner, then this person will have her job cut out. She has to explain, for example, why so many of us are systematically mistaken about what makes our lives go well. Thus, what we will say in chapters 4-6 about the ideology theory of oppression will still be relevant.

2.5 Cases of oppression

Although we can use the two criteria to determine whether a social order is unjust and thus whether it is oppressive, it is (for our purposes) even more important that people tend to dislike the social orders that violate these criteria. So far, we have argued that the criteria capture many of our intuitions concerning justice and that, therefore, people tend to dislike social orders that violate them. There are, however, also good reasons to believe that many actual revolutions and instances of civil unrest have been sparked by violations of these criteria.

Consider, for example, the public outrage against large bonuses in the financial sector. An unemployed manual worker who belongs to the group of the least advantaged in a social order may be willing to accept a social order where stock brokers receive million dollar bonuses if (a) this inequality is necessary to pay for her unemployment benefits and affordable healthcare, and (b) there are no formal barriers preventing her (or her children) from becoming stock brokers (and holders of other privileged positions). The reason why people are opposed to large bonuses in the financial sector, is that these bonuses do not have good consequences for, e.g., unemployed manual workers. In fact, it is often argued that the bonuses have very bad consequences since they give

stockbrokers incentives to take excessive risks. Much of the opposition to financial bonuses in the wake of the late-00s financial crisis was motivated by the fact that these bonuses had an adverse effect on the economy as a whole and consequently on the least advantaged.

Concerning the equality of opportunity condition, it could be argued that the unrest that lead up to the 2011 Arab Spring was caused in part by the fact that the positions associated with wealth and power were not open to all. A large part of the population was prevented from reaching positions of high economic and social status by widespread nepotism and corruption favouring family and clan members. Similarly, the 1960s protests of the American civil rights movement were clearly motivated by the fact that many positions in American society were not open to the African American population.

Equipped with a set of criteria that most people (under normal circumstances) tend to accept, we now turn to the task of identifying some persistent oppressive social orders. Our aim is to describe social orders (or aspects thereof, to be more precise) in the real world that satisfy conditions 1-3. Since it is an empirical matter whether a social order satisfies these conditions, we will point to different features of these social orders to support our claims.

2.5.1 North Korea

Although little is known to outsiders, it seems safe to say that the ordinary North Korean citizen suffers badly under the present regime. For example, North Korea's estimated GDP per capita (adjusted for purchasing power parity) is $1800, ranking it as the 160th largest economy in the world out of 193.[1] The output can also be compared to South Korea's $30,000. GDP per capita is of course not the only relevant measure of welfare in a country but it can be used as an index of the level of economic development. A more relevant index would perhaps be the United Nation's *Human Development Index* (HDI) that in addition to GDP per capita takes into account, e.g., gender equality, life expectancy, and literacy rate. Unfortunately, there is not enough information about North Korea to calculate HDI. There is, however, an estimate of the life expectancy at birth that can give us an indication of how healthy North Koreans are in comparison with, e.g., South Koreans. In North Korea, estimated life expectancy at birth is slightly under 67 years, resulting in a country rank of 125 out of 193. This can be compared to South Korea where it is slightly above 79 years. When it comes to basic rights, such as freely expressing one's opinion, North Korea ends up as the second worst country in the world on

[1] Unless explicitly stated the statistics about North and South Korea are taken from CIA's *The World Factbook* [2012].

Reporters Without Borders' *Press Freedom Index of 2010* with only Eritrea doing worse.[1]

It should, however, be pointed out that on some counts North Korea seems to perform relatively well. According to one report, prepared by the Bertelsmann Foundation, North Korea is doing relatively well when it comes to equality. For example, the government provides free education (including university) and health service for all. It could perhaps be argued that North Korea provides fair equality of opportunity for its citizens to attain some of the positions within its social order. It is, however, unlikely that everyone has equal opportunity to reach the extremely privileged positions within the ruling party.

Furthermore, we would be hard-pressed to claim that North Korea is doing well when it comes to promoting its citizens' welfare in terms of wealth and income, health, and freedom of speech. North Korea seems to be a paradigmatic example of a social order that does not satisfy the maximin criteria. It is also difficult to see any reason why North Korea could not have done much better had it had different political leadership and socioeconomic institutions more similar to those of, e.g., South Korea. It seems safe to claim that the citizens of North Korea are systematically treated unjustly, and that this therefore is an example of an oppressive social order.

Another feature of North Korea that makes it interesting for our candidate theories is that the citizens of its southern neighbour, South Korea, speak the same language and have a similar cultural history as the North Koreans. This was also true of the citizens of East German who had much in common with the West Germans across the border. These similarities allowed the citizens of East Germany (and allow the citizens of North Korea) to be relatively well informed about the living conditions on the other side of the border. Even if we disregard explicit propaganda in the form of, e.g., Radio Liberty broadcasts, there were a number of ways for the citizens of East Germany to get information about life on the other side of the Iron Curtain. They could, for example, tune their TV sets to receive West German programs, and read newspapers and magazines that had been smuggled across the border. In the case of North Korea, South Korean newspapers and magazines are being smuggled across the border from China. Furthermore, the demand for South Korean TV shows has created a second hand market for VCR players that are also smuggled across the Chinese border along with video recordings of South Korean TV shows. Sharing language and culture with one's closest neighbour allows oppressed citizens to be relatively well-informed of what life could be like in a less oppressive social order.

A second interesting feature that North Korea shares with East Germany

[1] http://en.rsf.org/press-freedom-index-2010,1034.html accessed on July 12, 2011.

(and the other members of the former Eastern Bloc and pre-revolution North Africa) is that the oppressed citizens vastly outnumbers the oppressive regime's ruling elite. According to one study, North Korea's ruling elite numbers less than 100 persons and governs more than 20 million (in 1993).[1] Although there are probably more who gain from and support the present regime, the number gives us an indication of the proportion of oppressors to oppressed. Given that they vastly outnumber their oppressors, it seems as the North Korean citizens would have little trouble overpowering their leaders.

North Korea is not the only regime where a very small number of people have ruled an entire nation, but it is clearly one of the most resilient. Whereas the East German Socialist Party's rule lasted for 41 years and Mubarak lasted for 30 in Egypt, the Korean Worker's Party has prevailed since 1948 and survived the deaths of "Beloved Leader" Kim Il-Sung and "Dear Leader" Kim Il-Jong, as well as the mid-90s famine.

North Korea is in many ways the perfect example of a persistent social order that satisfies condition 1-3. The citizens are oppressed since neither the fair equality of opportunity condition nor the maximin condition are satisfied. We also have good reason to believe that the North Koreans could have a much higher degree of welfare in a less oppressive social order. Finally, the North Korean people have the ability to exchange the present social order for a less oppressive social order since they vastly outnumber their rulers.

Other examples of oppressive social orders that have much in common with North Korea are South Africa under the apartheid era, present day Belarus, and, as we have already mentioned, East Germany and many of the other former countries of the Eastern Bloc. When we attempt to explain the persistence of tyranny in chapter 3 we will mainly be concerned with these social orders.

2.5.2 The American dream

Claiming that North Korea is an example of an oppressive social order is hardly controversial. Our next examples may, however, be a bit more controversial. We will now be concerned with relatively well-functioning Western democracies where the citizens have the ability to influence their social orders by participating in the democratic process. They can do this not only by voting and running for office in regular elections, but also by making their voices heard in public without (normally) having to fear violent reprisals. Therefore, it is more or less obvious that condition 3 is satisfied in these social orders. However, what is controversial is whether there are any groups that are oppressed in these Western democracies. In this section and the next, we will try

[1] See Andrea Savada [1993].

to show that some Western societies are indeed oppressive.

Let us begin by noting that everyone is not equally well off in the US. According to a report by NYU economist Edward Wolff [2010, p. 11], in 2007, the richest 20% owned 85%, and the bottom 40% owned 0.2% of the total wealth in the US.[1] This can be compared to Canada where, in 2005, the richest 20% owned 69.2% and the poorest 40% owned 2.4% of the total wealth.[2] Although the level of inequality is high in Canada, it is obviously lower than in the US. Furthermore, since Canada is culturally and geographically similar to the US it is difficult to see why the social order in the US could not be replaced by a less unequal Canadian-style social order. Therefore, it seems as the poorest American could be better off in an alternative social order.

That wealth is unequally distributed might not be a problem if the privileged positions in a society are open to all under fair equality of opportunity. The bottom 40% might accept that they own 0.2% of the total wealth if they could have belonged to (and their children have the chance of belonging) to the richest 20%. However, according to a 2010 OECD report [2010, ch. 5] comparing the OECD members with respect to intergenerational social mobility this does not seem to be the case.[3] It reports that in the US at least 40% of the economic advantage that high-earning fathers have over low-earning fathers is transmitted to their sons. This can be compared to the Nordic countries, Australia, and Canada where less than 20% of earning power is transferred from father to son. In other words, a high degree of social mobility cannot be used to explain the lack of social unrest.

Furthermore, the fact that income inequalities are persistent might give us reason to believe that the advantaged positions in society are not open to all under fair equality of opportunity. Although there are no explicitly discriminating laws in the US, there seem to be laws that are biased against certain disadvantaged groups. For example, the system where education is mainly financed by property tax seems to perpetuate inequalities since rich neighbourhoods tend to have a larger tax base than poor neighbourhoods. So even if we would not go as far as to demand free education and healthcare for all, there are still reasons to believe that the present system is biased against already disadvantaged groups.

It is also worth mentioning the findings of a poll performed by New York

[1] And the richest 1% owned 34.3%.

[2] See Statistics Canada [2005, p. 9].

[3] Intergenerational social mobility is measured as the correlation between the income of father and son. The higher the correlation, the more of the father's earning power is transferred to his son, and therefore the lower the intergenerational social mobility.

Times in 2005.[1] Although social mobility had dropped in the US since 1975, the majority of the respondents answered that they believed that there was more upward social mobility in 2005 than 30 years before. Of the respondents 40% answered that social mobility had increased, and only 23% answered that it had decreased. When asked to compare upward social mobility in the US with European countries, 46% answered that it was higher in the US. Of the European countries the OECD report compares, only Italy and the UK had a slightly lower level of intergenerational social mobility than the US.

Let us conclude that the persistent inequality of wealth and low social mobility in the US provides us with an example of the mystery of oppression. The poor Americans would be better off under a more equal social order. Furthermore, there does not seem to be any redeeming qualities in terms of high social mobility. Finally, being a relatively well-functioning democracy there are good reasons to believe that the poorest Americans could exchange the present social order for another, less oppressive, social order.

It is also interesting to repeat that although Americans do not perceive their society as oppressive or unjust, it is likely that Americans as most others dislike social orders with low social mobility and extreme inequality.

2.5.3 The gender wage gap

It has often been claimed that women are oppressed not only in countries employing obviously discriminatory laws, such as Saudi Arabia, but also in present-day Western democracies. In this section, we will show that there are good reasons to believe that women are indeed oppressed in Europe and the US.

The common claim that women earn less than men is easy to confirm. Using data from the U.S. Census Bureau on 2005 median earnings of men and women we see that women earn approximately 68% of what men earn.[2] The corresponding figures are for Sweden in 2009 85%,[3] and for the UK in 2010 80.2%.[4] Although Sweden and the UK fare better than the US, it is disconcerting that women earn less than men in three of the world's most developed democracies. The gender wage gap gives us reason to believe that the privi-

[1] See http://www.nytimes.com/pages/national/class/ accessed May 19 2011.

[2] This is calculated by using data for females 25 years and over from all races and males 25 years and over from all races from http://pubdb3.census.gov/macro/032006/perinc/new03_000.htm accessed on July 20, 2011.

[3] http://www.scb.se/Pages/TableAndChart____149083.aspx accessed on July 20, 2011.

[4] http://www.statistics.gov.uk/cci/nugget.asp?id=167 accessed on July 21, 2011.

leged positions within these social orders are not open to all under fair equality of opportunity.

It might, however, be claimed that although it is true that women earn less than men, this is not due to discrimination against women, but rather because women and men make different so-called life choices. For example, Kingsley Browne [2002, p. 73] argues that if we take into account that women value flexible working hours more than men and therefore tend to work less hours than men, the gender wage gap decreases drastically. Browne goes on to point to other factors that can explain the difference in pay that does not presuppose discrimination against women. He does, for example, point out that women tend to have a lower level of job-related schooling (more prone to getting a Master's degrees in education than a degree in business) [p. 75], and that women are less prone to ask for higher wages than men [p. 84].

It is relatively easy to test Browne's first two claims by controlling for hours worked and level of education when calculating the gender wage gap. The data from the U.S. Census Bureau allows us to control for hours worked by comparing how much men and women in full-time employment earn. When these data are compared we find that women earn approximately 77% of what men earn.[1] In the UK the corresponding figure is 89.8%. Thus, it seems as part of the gender wage gap can be explained by the fact that men tend to work more hours than women. It should, however, be clear that it does not explain the whole gap. Unfortunately, neither the U.S. Census Bureau nor the UK Office for National Statistics provides a comparison of men and women's wages where education level has been controlled for. Statistics Sweden, on the other hand, report that Swedish women earn 93% of what Swedish men earn when hours worked, education level, sector, and age has been controlled for.[2] Thus, even after life choice factors have been controlled for, there is still a 7% difference that is unaccounted for.

Having established that there is a gender wage gap left unexplained by life choice factors, discrimination may hold some explanatory power. It should also be pointed out that even if it was shown that the gender wage gap could be completely explained by differences in hours worked, career and educational choices between men and women, and so on, this would not mean that

[1]This is calculated by using data for females 25 years and over from all races and who worked full time all year round, and males 25 years and over from all races who worked full time all year round from http://pubdb3.census.gov/macro/032006/perinc/new03_000.htm accessed on July 20, 2011.

[2]Controlling for sector is relevant if women tend to work in sectors with lower wages than other sectors. Age is relevant if it is correlated with wage and women in the labour market tend to be younger.

the distribution is just or that it was not caused by discrimination. It could, after all, be the case that the life choices were influenced by unjust or discriminatory practices and legislation. For example, if a mother's choice to work flexible hours is caused by labour legislation that does not grant parental leave to fathers, then the labour legislation seems to be biased against women.

Furthermore, even if biology was the underlying cause of the gender wage gap (by priming women to choose family over career), as Browne seems to argue, it is not entirely clear that this would absolve the social order from its responsibility. At least not if people tend to accept some wider meaning of fair equality of opportunity that demand that a social order should see to it that individuals have the same prospect of success no matter their initial position in society or biological lottery. However, we do not need to dwell on whether biological differences can explain the gender wage gap, or whether this fact will remove the responsibility of social orders. We will simply assume that the biological differences between men and women are not large enough to explain the gender wage gap. Therefore, the gender wage gap gives us an example of oppression of women in Western democracies, and just as the previous example there should not be any question of whether women have the ability to exchange the present social order for another, less oppressive, social order.

One last feature of oppression of women that we will get reason to look at more closely in chapter 4 is connected to Browne's claim that women are less prone to ask for higher wages than men. Similar claims have been investigated in experimental situations. For example, Håkan Holm [2000] has performed an experiment with a group of Swedish undergraduates who were paired and asked to play a so-called clash of wills game.[1] The payoffs (in SEK) for the game is given by figure 2.1.

		B	
		Hawk	Dove
A	Hawk	$0,0$	$200,100$
	Dove	$100,200$	$0,0$

Figure 2.1: Clash of wills, payoff given in SEK.

The experimenters provided the players with information about their co-player's gender with the help of a generic Swedish name. For example, Karl Andersson, if the co-player was male, and Lisa Svenson, if the co-player was female.[2] The result of the experiment was that when a man was paired with a woman, men had a tendency to play hawk (77.1%) and women dove (35.7%),

[1] Sometimes called a battle of the sexes game.

[2] In order to avoid making the gender-aspect of the experiment too salient a generic

when two men were paired they tended to play hawk approximately half of the time (55.3%). If, however, they were allowed to split equally, the equal split was almost always chosen (94%). Holm draws the conclusion that when the equal split option is unavailable, a gender-based focal point biased towards men takes priority. The problem, for our purposes, is that the outcome is not obviously unjust. Women, as well as men, are after all made better off by the existence of a gender-based focal point in the sense that if it did not exist, everyone would be worse off.

Although the outcome of this experiment fails to show that women are systematically treated unjustly, it still provides us with some experimental evidence for the claim that women tend to give in to men in bargaining. It is, for example, often claimed that women in heterosexual relationships spend considerable more time on household chores than men.[1] The experiment might also give us an indication of where to look for micro-foundations for such claims. We will have reason to return to this in chapter 4.

For now, let us recapitulate. The gender wage gap seems to indicate that women are oppressed since it shows that the privileged positions (in terms of wealth) are not open to women under fair equality of opportunity. Furthermore, although the last experiment indicates the possibility that women are made better off because of gender-based focal points in some situations, it seems reasonable to assume that in the normal run of things being paid less than their male counterparts goes against the interest of women. Finally, it seems as women have the ability (in the same sense as the American poor) to exchange the oppressive social order for a less oppressive social order.

2.6 Summary

In this chapter, we have explicated the concepts of oppression and injustice. We argued that to call a social order oppressive is to imply that some group is treated unjustly. In order to avoid becoming committed to any specific normative or meta-ethical theory we argued that instead of talking about injustice, we should try to find some criteria that allows us to identify the social orders that people tend to consider unjust and oppressive. After all, for our purposes it is sufficient that people tend to dislike, or would under certain ideal epistemic circumstances dislike, the social orders we call oppressive. We suggested the following two criteria:

name was used to signal gender.

[1] See, e.g., Thompson and Walker [1989].

1. Social order *A* satisfies the THE MAXIMIN CRITERION if there is no other feasible social order, *B*, where the worst off individual in *B* has a higher level of expected welfare than the worst off individual in *A*.

2. Social order *A* satisfies the FAIR EQUALITY OF OPPORTUNITY if the social and economic inequalities in *A* are arranged so that they are attached to positions and offices that nobody is prevented by law (directly or indirectly through a systemic bias) from attaining.

Once we had established the criteria, we moved on to identify some persistent oppressive social orders. The persistence of the social orders described in section 2.5.1-2.5.3 is what our theories of oppression should be able to explain. With each example we tried to highlight some different aspects of persistent oppression that the theories should try to account for.

We began by describing the rather uncontroversial case of North Korea where we focused on the fact that a vastly outnumbered minority use violent repression to uphold an oppressive system. We then moved on to oppression in Western democracies. Our first stop was the USA. We argued that American society is oppressive in the sense that it is extremely unequal and has a very low level of intergenerational social mobility. Since the USA is a relatively well-functioning democracy it seems as America's poor could use democratic means to exchange the oppressive social order for a less oppressive social order. A theory of persistent oppression should, therefore, be able to explain why poor Americans do not march against Washington DC demanding reforms. With the risk of loading the dice in favour of ideology theory of oppression, we pointed out that Americans seem to be unaware of the relatively low level of intergenerational social mobility in their society.

For our final example we looked at the differences in wages between men and women in Sweden, the UK, and the US. We showed that after we have controlled for life choice factors there is still a part of the wage difference between men and women that remains unexplained. We claimed that the so-called gender wage gap indicates that women are oppressed in Western democracies. That the members of one group systematically get paid less than the members of another group, when the only difference between the two groups is with regard to reproductive organs, is indeed something that a theory of persistent oppression should explain.

Armed with three examples of persistent oppressive social orders, we can move on to evaluate the claim that the gunman theory of oppression is able to explain these cases. In the next chapter we will show that the gunman theory of oppression does a very good job at explaining persistent tyranny in the oppressive social orders found in North Korea, East Germany and Mubarak's Egypt. We will then, in chapter 4, show that it does worse when it comes to

explaining the persistent economic inequality in the US and the gender wage gap in Sweden. We will also show that the ideology theory of oppression is needed to account for the aspects that make these cases interesting.

3. Rational revolutionaries

3.1 Introduction

One of the purposes of Alexis de Tocqueville's and Gustave de Beaumont's 1831 journey to America was to study the American prison system and its applicability in France. The study resulted in a book aptly entitled *On the penitentiary system in the United States and its application in France*.[1] Among other things, the book contains an explanation of the stability of the tyrannical structure at the Sing-Sing prison.

The phenomenon they stumble upon at Sing-Sing is described in a letter from Beaumont to his mother:

> It is certain that the system of discipline established in the penitentiary of Sing-Sing is very remarkable [...] This prison contains 900 inmates [and 30 guards]. They [the inmates] are at complete liberty, carrying irons neither at hand nor feet, and yet they labor assiduously at the hardest tasks. Nothing is rarer than an evasion. That appears so unbelievable that one sees the fact a long time without being able to explain it. [Beaumont cited in Pierson [1959, p. 67]]

The depicted scene has a lot in common with some persistent oppressive social orders. The same remark Beaumont and Tocqueville [1833, p. 26] make about Sing-Sing can be made about many oppressive social orders: "[i]t is evident that the life of the keepers would be at the mercy of the prisoners, if material force were sufficient for the latter." If it came down to an all-out fight between the 30 guards and the 900 prisoners, then the prisoners would surely win. The same is probably true for the relationship between the ruling elite and the oppressed people in, e.g., North Korea, Iran, and Belarus. If all members of the oppressed majority would take to the streets simultaneously, the oppressive minorities would not stand a chance. Furthermore, it is likely that just as the prisoners detest labouring for the guards, the citizens of North Korea, Iran, and Belarus would prefer to exchange the present social order for a less repressive order. Thus, it is surprising that the prisoners work for the guards for much the

[1] Beaumont and Tocqueville [1833].

same reasons as a majority keeps labouring for an oppressive minority year after year.

Now, Beaumont's and Tocqueville's explanation of how a few guards could keep a much larger group in check was the following:

> [b]ecause the keepers communicate freely with each other, act in concert, and have all the power of association; whilst the convicts separated from each other, by silence, have, in spite of numerical force, all the weakness of isolation. [Beaumont and Tocqueville, 1833, p. 26]

In other words, the guards' ability to communicate allowed them to coordinate their use of force to isolate the prisoners and prevent them from communicating. Consequently, it was as if each and every one of the 900 prisoners individually faced a 30-man strong group of guards.

With the similarities between Sing-Sing and tyrannical oppressive social orders in mind, the same explanation might be used to explain the persistence of tyranny. To be more precise, Tocqueville's argument suggests that the oppressive regime can take precautions that will prevent the oppressed from getting rid of them. For example, oppressive social orders can use press censorship and inhibit freedom of speech to prevent the oppressed from coordinating their efforts. Similar theories of persistent oppression have been advocated by, for example, Russell Hardin [1995], Michael Rosen [1996], and Magnus Jiborn [2006]. As we will see, one of the great merits of a gunman theory of oppression that incorporates the power of communication is that it can avoid the objection that the gunman theory cannot explain why oppressive social orders sometimes break down.

The gunman theory of oppression is often formulated within an elegant game theoretic framework. There are, however, some problems involved in using game theory to account for the role of communication in oppression. It has, for example, been claimed that if rationality is so-called common knowledge then, in the kind of games that are used to model the oppressive situation, rational agents will not be swayed by any kind of communication. We will show that this problem can be avoided by replacing the common knowledge assumption with some more realistic assumptions about the beliefs of the involved agents.

The purpose of this chapter is to formulate a plausible version of the gunman theory of oppression and show that it does a good job explaining persistent tyrannies in countries such as North Korea, East Germany, and Belarus. In the next chapter, we will show that although it can explain a great deal about persisting oppression it cannot explain all persistent oppressive social orders. Thus, we will have a reason for investigating whether the ideology theory of

oppression can be demystified and used to complete the explanation of oppression.

The chapter will be organised as follows. In section 3.2, we will go through some assumptions and introduce the definitions that will be used in formulating the explanation. Next, in section 3.3, we will formulate the game theoretic problem that face the prisoners of Sing-Sing as well as oppressed citizens. We will see that under simple assumptions about rationality there is only one possible outcome of this game: nobody revolts and nobody attempts to escape. This might seem as good news (for the gunman theory of oppression) if it was not for the fact that at times both prison riots and revolutions break out. In section 3.4, we will modify the assumptions describing both the Sing-Sing prison and the revolutionary situation so that they become somewhat more realistic. Under the new assumptions both stability and a revolt can be the outcome of rational action. The problem is then to explain why stability breaks down when it does. We will see that although Beaumont's and Tocqueville's silence-account seems intuitive enough, communication (or the lack thereof) has no effect on the actions of agents for whom rationality and the structure of the game are common knowledge. We will therefore show that if the common knowledge assumption is replaced by a, so-called, level-κ model of strategic thinking it is possible to account for the role of communication.

Although the level-κ model provides a more realistic description of the formation of beliefs than the common knowledge assumption, it is still obviously false (at least if taken literary). In section 3.5, we will address the problem of providing explanations with false assumptions. Section 3.6 concludes.

3.2 Definitions and assumptions

The gunman theory of oppression is founded on the theory of rational choice. Rational choice theory provides a formal framework for describing and understanding how rational agents make decisions. The theory's starting point is the Humean theory of action that holds that both a pro-attitude and a belief are necessary in order to produce an action. More specifically, it can be said to rest on the following 'folk psychological' law: if an agent, A, wants d, believes that a will bring about d, and can do a, then A will do a. It is, however, easy to find exceptions to this claim. For example, A might refrain from doing a if she has a stronger desire, d', that can only be satisfied by not doing a. Therefore, we cannot use it as the nomological component of an explanation that complies to the deductive-nomological model.

Rational choice theory provides a specification of folk psychology that

can be used to explain actions with the help of pro-attitudes and beliefs.[1] It begins by assuming that when a rational agent faces a decision problem that forces her to choose between a number of alternatives, she is be able to rank all alternatives with respect to how well they satisfy her preferences. More formally, it is assumed that for all possible pairs of alternatives, x, y, the rational agent will prefer x to y, or prefer y to x, or be indifferent between them.[2] This is sometimes called the assumption of *completeness* since it implies that rational agents have a complete ranking of all available alternatives. For example, a rational agent who walks into a supermarket in order to buy fruit is assumed to have a complete preference ordering over all the fruit.[3]

The theory also assumes that rational agents have *transitive* rankings. Formally, it is assumed that for any three alternatives, x, y, z, if an agent prefers x to y, and y to z, then she prefers x to z.[4] This means, that if she prefers apples to oranges, and oranges to pears, then she must prefer apples to pears.

Given the completeness and transitivity assumptions there will always be a set of alternatives that maximise the preference satisfaction of a rational agent in every decision problem. Rational choice theory then makes a third assumption: a rational agent is assumed to always choose one of the available alternatives that maximises her preference satisfaction.

Note that the assumptions so far have been about rational agents. Since it is possible that there are no rational agents in the real world, something is needed in order to connect the theory of rational agents to the world of real agents. A fourth assumption takes care of this: people are rational in the sense that they have complete, transitive preference orderings, and attempt to maximise their preference satisfaction. If the assumptions are correct and if we know the preference orderings of the agents in a given choice situation, then it is possible to use rational choice theory to explain and predict how they will act in this situation.

However, besides requiring transitivity and completeness rational choice theory does not say anything about the preference orderings of rational agents. Therefore, there is nothing irrational about favouring immediate preference satisfaction to future satisfaction, or satisfying ill-informed, or non-autonomous

[1]For an introduction on rational choice theory see, e.g., Alexander Rosenberg [2008, pp. 80-5] and Michael Resnik [1987, ch. 2].

[2]However, see, e.g., Ruth Chang [2002] for a critique of this assumption.

[3]Usually she is not only assumed to have a preference ordering over the types of fruit, e.g., apples over oranges, but also over the all fruit-tokens, e.g., apple_1, over orange_1, over apple_2, etc.

[4]The transitivity assumption also covers cases where the agents is indifferent between alternatives. A rational agent who is indifferent between x and y and prefers y to z, also prefers x to z.

preferences. The basic idea is famously expressed by David Hume [1985, p. 463]: "'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger."

It should also be pointed out that, contrary to common belief, rational choice theory does not imply that rational agents never care about the welfare or preferences of others. Once again, in the words of Hume [1985, p. 463]: "'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an Indian or person wholly unknown to me."

Many applications of rational choice theory do, however, add the assumption that rational agents are relatively self-regarding in the sense that they care relatively more about what happens to themselves than what happens to others.[1] Since it allows us to ignore potential other-regarding motivations this assumption makes it easier to derive predictions of individual behaviour. In the case potential revolutionaries, it allows us to assume that individual *i* does not care about *j*'s welfare when *i* decides whether she should participate in a revolt.

Finally, rationality does not even require agents to care about their own welfare. For example, the loyal work horse from George Orwell's *Animal Farm*, Boxer, who unconditionally supports Napoleon's regime and works for the common good with no regard for his own health or welfare does not count as irrational as long as he satisfies the above conditions.

In order to make meaningful claims and prediction, however, more substantial claims about the preferences of rational agents are needed. This is where the gunman theory of oppression enters the picture. With respect to our specification of the mystery of oppression in chapter 2, the gunman theory assumes that the oppressed believe that they are oppressed and that they are motivated to maximise their own welfare (or interest-satisfaction). It is, in other words, assumed that they have correct beliefs about their situation, and that their desires and interests coincide. Furthermore, all oppressed are assumed to know the risks they take by participating in a revolt, the chances of success, and the value (in terms of welfare) of a life in a less oppressive social order.

The substantial assumptions of the gunman theory of oppression can be questioned. It is far from obvious that all oppressed are motivated by a desire for a less oppressive social order. For example, it is possible that some believe that they ought to be oppressed, and therefore prefer to stay in the oppressive social order. It is also far from obvious that rational agents care more about themselves than they care about others. There is, after all, plenty of evidence that people are willing to sacrifice their own welfare for the good of others.[2]

[1] See, e.g., Philip Pettit [1995, p. 310].

[2] See, e.g., the experimental work of Ernst Fehr and Klaus Schmidt [1999].

We will offer a defence of the use of unrealistic assumptions in section 3.5. For now, however, let us accept these assumptions and proceed to have a look at the more specific assumptions about the tyrannical social order.

The assumptions of the gunman theory of oppression can be applied to an idealised version of a tyrannical social order (e.g., North Korea). First, let us assume that there are two groups of people: a large group of *n oppressed* and smaller group of *oppressors*. The oppressors want to maintain the social order, whereas the oppressed (as we mentioned above) all share a desire to replace the present social order with a less oppressive social order.

Next, assume that the only way the oppressed can get rid of the oppressive social order is by staging a successful *revolution*. If at least one of the oppressed decides to *participate* in a revolution, then we say that a revolutions breaks out; and if a certain number, $1 \leq k \leq n$, of oppressed participate, then the revolution will be successful. The relevant choices of the oppressed will thus be to either participate in a revolution, or to remain *inactive*. Furthermore, if a revolution succeeds then all oppressed will be allowed to escape, whether they chose to participate in, or abstain from the revolutionary activities.

The oppressors, on the other hand, will not be considered players and, therefore, will not be given any choices. They will simply be assumed to attempt to punish everyone who participates, but never those who abstains. It will also be assumed that they will keep on punishing participants even if it is certain that the revolt will succeed. Since the oppressors are assumed to follow this simple script, we will, in what follows, focus on the decisions of the oppressed and take the behaviour of the oppressors as given.

Becoming a bit more formal, let the structure of the game be given by the triple $G =<$Players, Strategies, Payoff$>$. In our case, the *players* are the *n* oppressed, the strategies are *participate* or remain *inactive*, and the *payoff function* specifies the welfare each player gets as a function of her own and all other players' strategy choice.[1]

Now that we know the structure of the game, let us introduce an assumption often made in game theory. It is assumed that the rationality of the players and the structure of the game are *common knowledge* among the rational agents. If we let *R* be the claim that all oppressed are rational, then we can formulate the

[1]Formally, a game is a triple $G =< N,S,U >$ such that $N = \{1,2,...,n\}$, and $S = \{S_1,S_2,...,S_n\}$ where $S_i = \{1,2,...,m_i\}$ is the set of pure strategies belonging to player i. Let $s_i \in S_i$ be the pure strategy played by player i, and $s_{-i} = (s_1,s_2,...,s_{i-1},s_{i+1},...,s_n)$ be the pure strategies played by all other players except i. We can then represent a pure strategy profile as $s = (s_i,s_{-i})$. The payoff function, $U : S_1 \times S_2 \times ... \times S_n \mapsto \mathbb{R}$, ascribes a payoff in real numbers to each player for all possible pure strategy profiles. For example, the payoff to player $i \in N$ if the pure strategy profile (s_i,s_{-i}) is played, is $u_i = u_i(s_i,s_{-i})$.

common knowledge claim as follows: R and G are common knowledge among the oppressed if

1. all oppressed know that R and G,

2. all oppressed know that 1 is known by all oppressed,

...

m. all oppressed know that $m-1$ is know by all oppressed,

where m approaches infinity.

It is often objected that the common knowledge assumption is far too strong, and that there is plenty of evidence that common knowledge does not obtain in games involving real people. It is, after all, enough that one of the m propositions is false in order for common knowledge to break down. One way around this is to say that it is enough that people behave *as if* the common knowledge assumption was true.

The problem is, however, that people do not even seem to behave *as if* common knowledge obtained. This can be demonstrated by the discrepancy between how ideally rational agents are predicted to play under common knowledge and how real people play the so-called Keynesian Beauty Contest (KBC). The KBC is a game where a group of players are simultaneously told to write a number between 1 and 100 and told that whoever writes the number closest to $2/3$'s of the group's mean will win a price. They are given this information simultaneously in front of each other in order to ensure that everyone knows that everyone knows that ... everyone has received this information. The unique solution to the game, if played by ideally rational agents where common knowledge obtains, is to write 1.[1] Real people, however, do seldom write 1 the first time they play the game; and if they would happen to write 1,

[1]To see this, imagine that we have a large group of ideally rational players where rationality and structure of the game is common knowledge. Let us ask what number, s_A, player A should write down in this situation if she wants to win. If everyone else writes 100 she will win by writing the integer closest to $\frac{2}{3}100$, i.e., $s_A = 67$. Therefore, she can immediately eliminate all $s_A > 67$. However, since she knows that everyone else is rational, and that everyone else knows that everyone is rational, and that everyone knows that everyone knows that everyone is rational, etc., she understands that nobody will write a number greater than 67. She will then realise that if everyone else would write 67, then she would win by writing the integer closest to $\frac{2}{3}67$, i.e., $s_A = 45$. However, for the same reasons as above she understands that nobody will write a number greater than 45. This iterated elimination of dominated strategies will go on until there is only one strategy left, $s_A = 1$. Since nobody will be willing to unilaterally change her number if everyone have written 1, this will be the unique equilibrium solution to the game. We will return to the concept of an equilibrium solutions below.

they seldom win the prize since a substantial part of the other players write a number larger than 1.[1]

If we want to use game theory to explain real world phenomena, then we seem to have a good reason to abandon the common knowledge assumption. This is something we will return to below. For now, let us just point out that it will be shown that in order to make sense of communication the gunman theory of oppression will have to abandon the common knowledge assumption.

3.3 A free-rider problem

Given the description of the revolutionary situation and the assumptions of rationality, it is easy to show that rational revolutionaries are subject to a free-rider problem. Think of Che, the rational revolutionary, who contemplates whether he should participate in an upcoming revolution. He knows that if everyone else participates, then the revolution will succeed with or without his help, and since he might get punished if he participates he will be better off remaining inactive. Furthermore, he knows that if everyone else abstains from revolutionary action, then the revolt will fail whether he participates or not. And since the risk that he will be punished is great if he is the only one who challenges the regime, there is no point in him participating by himself. If it is assumed that everyone is as rational as Che, and common knowledge obtains among the oppressed, then it is clear that everyone would reason as Che, and thus nobody would participate in a potential revolution.

Allen Buchanan [1979] employs a similar argument to describe the proletarian revolution as a two-player game between the individual proletarian and the other proletarians. In Buchanan's model, the game is played by the individual revolutionary against all others. Both the individual and the others are assumed to choose between participating and remaining inactive. The payoffs for the individual are represented in figure 3.1.

		Others	
		Participate	Inactive
Individual	Participate	3	1
	Inactive	4	2

Figure 3.1: An individual revolutionary's payoff for participating or remaining inactive.

The worst thing that can happen to the individual revolutionary is that she contributes when the others remain inactive. She will then pay the cost for

[1] See, e.g., Rosemarie Nagel [1995].

contributing while at the same time not getting the benefit of a successful revolution. This is represented by the payoff 1. The next worst thing is if everyone, including her, remains inactive. Although she will not have to pay the cost of contributing, she will miss out on the benefits of a successful revolution. This outcome gives her a payoff of 2. To this she will prefer the outcome where everyone, including her, participates. Here she will pay for contributing and get the benefit of the revolution, for a payoff of 3. However, the best thing that can happen to her is if everyone else participates and she remains inactive. Getting the benefit of a successful revolution for free awards her a payoff of 4.

Since we have assumed that all potential revolutionaries are in the same situation, the payoff structure will be the same as the payoff structure of the so-called prisoner's dilemma. The prisoner's dilemma is usually presented as a story about two suspects who have been arrested by the police. The police officers do not have enough evidence to get the suspects convicted so they come up with a way of forcing a confession. They offer both suspects the following deal: if one of them confesses and the other remains silent, then he who confesses will walk free whereas the other gets ten years in prison. If both confess, then both get five years. If both remain silent, then the officers will charge them with some minor offence and they will get one year each.[1]

If we assume that the suspects only care about spending as little time in prison as possible, then figure 3.2 accurately describes their situation. The pay-

		Suspect 2	
		Silent	Confess
Suspect 1	Silent	$1,1$	$10,0$
	Confess	$0,10$	$5,5$

Figure 3.2: The prisoner's dilemma. Payoffs are years in prison.

offs associated with confessing strictly dominate the payoffs associated with remaining silent.[2] If suspect 2 remains silent, suspect 1 walks free by confessing; and if suspect 2 confesses, suspect 1 gets five years instead of ten if she

[1]The dilemma is nicely illustrated in the opening scene of season five of *The Wire* when a police officer attempts to make one of two suspects confess. The police officer tells the suspect that "You see now, I'm here to tell you this remainin' silent shit ain't nothin like they make it up to be. You up in here all tight with it waitin' for your paid lawyer. No, see, that work when you're some kinda criminal mastermind, when you haven't been seen runnin' from the deed. When you're own fuckin' runnin' partner ain't in the next room puttin' you in. Oh yeah. He's tellin' it like a bitch. We even went to Mickey D's for him because he was so motherfuckin' helpful. Two quarterpounders. Big fries. McDonaldland cookies. Dr. Pepper... That's how your boy roll, right?"

[2]Formally, a two-player prisoner's dilemma is a game, $G =< N,S,U >$, such

confesses. Thus, no matter what suspect 2 does, suspect 1 will be better off confessing. Since the same is true for suspect 2, and since both suspects are assumed to be rational and self-interested both will confess. They will do this although the outcome where both remain silent would be better for both.

In game theory, a set of strategies is said to be a *Nash equilibrium* if, and only if, none of the players would be better off by unilaterally changing her own strategy.[1] In the prisoner's dilemma game the only strategy pair that has this property is (confess, confess). What is interesting is that there is another outcome, (silent, silent), that would be a *Pareto improvement* compared to (confess, confess), since both players would be better off if this strategy pair was to be played. Thus, if the prisoners had the ability to enter into an agreement to remain silent *and* to enforce compliance, they would both be better off. However, barring such a possibility they will both confess.

We can describe the revolution as a two-player prisoner's dilemma if we assume that the population consists of Che and Fidel and that the revolution will succeed if at least one of them participates. If only one of them participates, then he who participates will become badly hurt and will not be able to enjoy his newly achieved liberty. If both participate they will both get seriously injured. The injuries will, however, not be severe enough to make them prefer a life under oppression to a life in liberty. Given these assumptions, we can represent Che and Fidel's situation with the two-player game shown in figure 3.3. The payoff structure of this game is identical in structure to a

		Fidel	
		Participate	Inactive
Che	Participate	3,3	1,4
	Inactive	4,1	2,2

Figure 3.3: Che and Fidel's dilemma. Payoffs represent welfare.

prisoner's dilemma. The payoffs for remaining inactive strictly dominate the payoffs for participating. Thus, no matter what Fidel decides to do, Che will get a strictly higher payoff by remaining inactive. The same is true for Fidel's payoffs. Since a rational and self-interested agent never plays a strategy that gives him a lower payoff than another, both will remain inactive.

that $N = \{1,2\}$ and $S_i = \{sil, con\}$ (*sil*=silent, *con*=confess) for all players $i \in N$, and with following payoff structure for all $i \in N$: $u_i(con_i, sil_{-i}) > u_i(sil_i, sil_{-i}) > u_i(con_i, con_{-i}) > u_i(sil_i, con_{-i})$. See also Magnus Jiborn [1999, p. 58].

[1] Formally, a strategy profile $s = (s_i, s_{-i})$ is a pure-strategy Nash equilibrium if, and only if,

$$\forall i \in N, s_i \in S_i, s_i \neq s_i^* : u_i(s_i^*, s_{-i}^*) \geq u_i(s_i, s_{-i}^*).$$

If we move from the two-player to the n-player case, we can turn to a more formal version of the revolutionary situation offered by Gordon Tullock [1971] and influenced by Mancur Olson's [1971] work on public good provision. Tullock argues that although revolutionary rhetoric often contains references to the common good, this good cannot be what motivates self-interested and rational revolutionaries. Tullock refers to the discrepancy between what revolutionaries say and what motivates them as the paradox of revolution.

Che who, once again, contemplates whether he should participate in an upcoming revolution can be used to illustrate Tullock's argument. This time, however, Che focuses on his marginal contribution to the success of the revolution. He estimates that without his contribution the probability that the revolution will succeed is p, and that it will be unsuccessful is $1-p$. Furthermore, he believes that his payoff is A if the revolution succeeds, and B if it is unsuccessful. Since we have assumed that all potential revolutionaries prefer a successful revolution to status quo, we have $A > B$. Furthermore, Che believes that if he participates in the revolution he will increase the probability of revolutionary success by p_p, and that he will subject himself to a risk of being punished at an expected cost of $C > 0$. That C is strictly greater than zero represents Che's belief that he can get hurt if he engages in revolutionary action. The expected payoff of his alternatives can be represented by the following two equations:

$$U_{\text{inactive}} = Ap + B(1-p), \tag{3.1}$$

$$U_{\text{participate}} = A(p+p_p) + B(1-(p+p_p)) - C. \tag{3.2}$$

Furthermore, if the number of participants is very large then it is unlikely that the revolution stands and falls with one individual. This seems to be the case in most revolutions that involve mass mobilisation. For example, if we had removed one person from the crowd in Independence Square in Kiev in 2004, or added one to the crowd in Tiananmen Square in 1989 the outcomes would hardly have been any different. In the model this is represented by setting $p_p = 0$, giving us the following approximation of (3.2):

$$U_{\text{participate}} \approx Ap + B(1-p) - C. \tag{3.3}$$

Since C is assumed to be strictly larger than zero the approximated expected payoff (3.3) for participation is strictly less than the payoff for remaining inactive (3.1). The result is thus the same as above: Che, the rational revolutionary, will not engage in revolutionary action if there are no other incentives than the common good to influence his decision.

Since A represents the value of a successful revolution to Che, and since A drops out of the equation if we subtract (3.1) from (3.3), Tullock concludes

that we can include any altruist consideration we can think of and still get the same result. Che might, for example, cry himself to sleep at night over the condition of the poor peasants. As equation (3.3) stands, he might be willing to sacrifice *almost* anything in order to improve their living conditions, and still refrain from revolutionary action since he does not think that his contribution will make any difference.

This is probably true of most real-life revolutionaries. It is, however, possible that a revolutionary is motivated by an ideal or altruistic consideration that outweighs the cost of participating no matter how small her own contribution to the success of the revolution. A convinced Kantian revolutionary would, for example, give up anything to satisfy the categorical (revolutionary) imperative. The same would be true of an extreme altruist. It is logically possible for an altruist, who greatly values the welfare of others, to get a high enough A to block the approximation of (3.2) to (3.3). A revolutionary will, after all, be willing to join the revolution if, and only if, the expected payoff for participating exceeds the expected payoff for remaining inactive, i.e., equations (3.2)-(3.1) ≥ 0:

$$A(p+p_p)+B(1-(p+p_p))-C-(Ap+B(1-p))\geq 0, \tag{3.4}$$

$$(A-B)p_p-C\geq 0. \tag{3.5}$$

In other words, if A is large then, in spite of p_p being very small and the risk of participating, C, large, the expected value of participating may exceed the expected value of remaining inactive.[1]

Although it is theoretically possible that revolutionaries are motivated by very strong altruistic considerations, this is probably not that common in real life revolutionary situations. Let us, therefore, accept Tullock's claim and assume that when Che's individual contribution to revolutionary success is very small and Che has some fear of being punished, he will not participate since he realises that his contribution makes no difference.

We can use a graph to illustrate the formal similarities between the n-player and two-player revolutionary game. In order to identify the Nash equilibrium of the n-player game we have to examine the expected payoff as a function of the number of participants. Let us, therefore, define the expected value, V, of

[1]Derek Parfit [1987, pp. 73-5] refers to the mistake of ignoring very small chances as 'the third mistakes in moral mathematics.' Parfit is, however, mainly concerned with what people *ought* to do, whereas Tullock is interested in what people actually do. In order for altruists to make a difference for the outcome of actual revolutionary situations, it is necessary that a significant number of potential revolutionaries are strong altruists. However, if revolutionaries are motivated by very strong altruistic considerations, then it could be argued that we have left the domain of the rational agents of the gunman theory of oppression and instead entered the domain of the ideology theory. For an opinion to the contrary, see Torbjörn Tännsjö [2007].

the revolution for an individual as a function of the number of participants, d, as

$$V(d) = Ap(d) + B(1 - p(d)). \tag{3.6}$$

Although Tullock argued that the individual contribution does not affect the probability of revolutionary success, it is still reasonable to assume that the probability of success increases with the number of participants. That is, although the addition or subtraction of some few revolutionaries does not seem to make any difference, the addition or subtraction of a large number probably affects the outcome. Had we, for example, halved the number of people at the Tahrir Square or doubled the number people at Tiananmen Square, then the outcome might have been different. Furthermore, it seems reasonable to assume that after some point each additional revolutionary contributes less to the expected value of the revolution than the previous added revolutionary. That is, the value of the revolution, $V(d)$, is subject to diminishing marginal returns with respect to the number of revolutionaries, d. Finally, if we assume that the risk of punishment, C, is constant, Thomas Schelling's [1978] representation of n-player games can be used to illustrate the revolutionary free-rider problem as a n-player prisoner's dilemma as in figure 3.4.

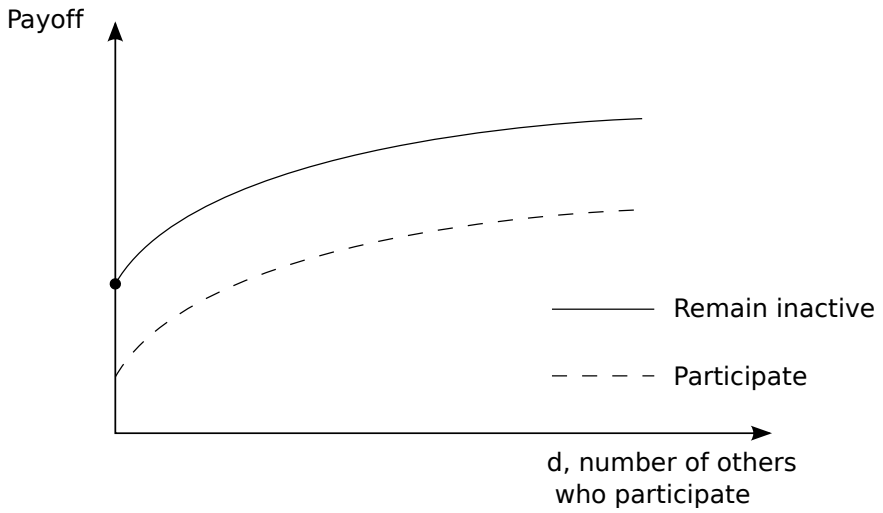

Figure 3.4: Payoffs with constant risk, the n-player prisoner's dilemma.

Just as in the two-player prisoner's dilemma the payoff for Che, the rational revolutionary, associated with remaining inactive is greater than the payoff for participating whatever the rest of the revolutionaries decide to do. Thus, if everyone is rational, then everyone will remain inactive no matter what. The black dot in figure 3.4 represents the unique Nash equilibrium in this situation.

We are now in a position to make the explanation offered by the gunman theory of oppression explicit. Recall that we formulated the mystery of oppression as follows:

> THE MYSTERY OF OPPRESSION: There exists a persistent social order, O_1, a possible persistent social order, $O_2 \neq O_1$, and a group X, such that
>
> 1. X is more oppressed in O_1 than in O_2,
> 2. O_2 is better than O_1 (in terms of welfare) for the members of X, and
> 3. X has the ability to exchange O_1 for O_2.

The gunman theory does not only accept 1 and 2, it also adds the assumptions that the revolutionaries believe that they are oppressed, and that they are motivated to maximise their own welfare.

In order to get rid of the mystery, the gunman theory focuses on claim 3 and argues that although the revolutionaries as a group have the ability to exchange the present social order for another, the free-rider problem prevents them from achieving this. We can follow Gregory Kavka [1986, p. 268] and summarise the argument as follows:

(a) Rational individuals act to maximise expected payoffs.

(b) In a potential revolution, the expected costs of participation are higher than the expected costs of non-participation, and there are no sufficiently compensating expected benefits of participation.

(c) Therefore, rational individuals in a potentially revolutionary situation will not participate in a revolution.

3.4 Overcoming the free-rider problem

The problem with the above model is that it is at odds with empirical findings. In the classical American, French, and Russian revolutions, and the more recent Ukrainian, Tunisian, and Egyptian revolutions, seemingly rational individuals both turned up in great numbers and successfully toppled the incumbent regimes. In other words, the empirical findings contradict the rational choice argument (a)-(c).

One way of resolving the contradiction is to claim that real people are not rational as described by premise (a). This is potentially bad news for the possibility of a rational choice explanation of stable oppression. For although

the gunman theory is able to explain why oppression is stable, it is unable to explain why it breaks down. This inability could open the door for other theories to explain why oppression breaks down. It could, for example, be argued that in the normal run of things citizens are rational and attempt to maximise their expected welfare. At times, however, ideology takes hold of them, offsets their rational considerations and allows them to stage a revolt. It could then be asked that if ideology is necessary to explain revolutions, why should it not be used to explain the persistence of oppression?

However, the proponents of rational choice theory avoids ideology by arguing that the second premise, describing the revolutionary situation, is inaccurate. Russell Hardin [1995, p. 52] has offered an explanation as to why rational individuals engage in revolutionary activity that consists of two components. The first springs from the observation that the risk of participation is not constant with respect to the number of participants. He observes that usually when the number of participants becomes large enough, the possibility of punishing a revolutionary starts to decline.

This observation is formalised by Magnus Jiborn [1999, p. 139] who models the state as a sanction system that is able to investigate and punish only a limited number of transgressions at a time. Consequently, when the number of transgressors exceeds the system's capacity, the risk of getting caught goes down. According to Jiborn's model the expected cost of sanctions, i.e., C, is inversely proportional to the number of participants once the capacity of the sanction system has been exceeded.

We can illustrate this with the help of the three rational revolutionaries Che, Fidel, and Raúl. They individually contemplate whether they should revolt against Fulgencio who has a sanction system at his disposal with the capacity to punish one revolutionary by inflicting a punishment of disvalue c. If Che is the only one who participates, he will get the disvalue c. If, however, Raúl participates as well, then there is a one-in-two chance that Raúl will get punished instead of Che, resulting in the expected disvalue $\frac{1}{2}c$. If all three participate, then there is only a one-in-three chance that Che gets punished, resulting in an expected disvalue $\frac{1}{3}c$.

The general equation for the expected disvalue of participating as a function of the number of participants in the n-player game can probably be made arbitrarily complicated. However, let us here assume, for simplicity's sake, that the capacity of the sanction system is one, and that the expected disvalue, C, is inversely proportional to the number of participants:

$$C(d) = c\frac{1}{d}. \tag{3.7}$$

Adding (3.7) to the expected value of the revolution (3.6), gives us the

following payoff for participating:

$$U_{\text{participate}} = V(d) - C(d). \tag{3.8}$$

The new revolutionary situation is shown in figure 3.5. The figure shows that

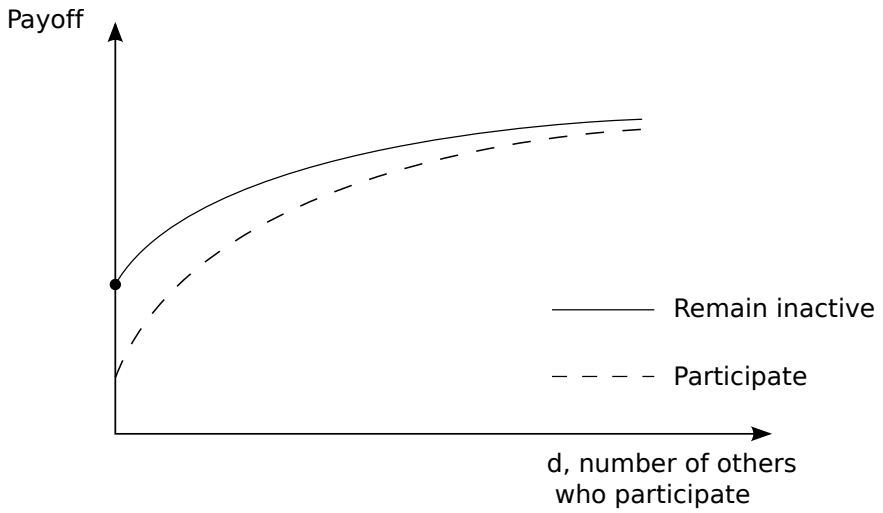

Figure 3.5: Payoffs with diminishing risk, still a *n*-player prisoner's dilemma.

the payoffs for participating and remaining inactive tend to converge as the number of participants increases. In other words, it becomes less profitable to remain inactive compared to participating as the number of participants increases. This can be used to explain why it becomes easier to attract followers to a revolution if the crowd is already large. Think of the Monday demonstrations in East Germany that started off as small-scale prayers for peace organised by the Leipzig Nikolai church, but later quickly spread across the country resulting in mass mobilisation. It could be argued that as the number of participants increased, the expected cost of participating became lower and thus induced more people to join in.

However, although it seems reasonable to assume that the expected cost diminishes as the number of participants increases, this argument fails to show how a successful revolution can be the outcome of rational action. The unique Nash equilibrium of the new game will still be that all remain inactive. This is once again represented by the black dot to the far left in figure 3.5. The rationale behind the result is that no matter how many revolutionaries take to the streets, there will still be some risk associated with getting involved compared to staying home and watching the events unfold on TV.[1] At best, Hardin's argument so far can show that as the number of participants tends to infinity the

[1]Because, contrary to Gil Scott-Heron, the revolution will most probably be televised.

expected cost for each individual revolutionary approaches zero, thus making the potential revolutionary indifferent between participating and remaining inactive. However, if we believe that there are some additional private costs that cannot be shared with others, such as the effort of getting up from the couch, then the payoff for remaining inactive will still dominate the payoff for participating.

This last observation brings us to the second component of a rational choice explanation of revolutionary action among rational revolutionaries. Although Tullock claimed that the common good does not motivate revolutionary action, he does not hold that revolutionary action is never rational. Rather, he follows Olson [1971] in arguing that collective action is possible due to the existence of so-called 'by-products'. According to Olson, In order to get people to overcome their incentives to free-ride, selective incentives have to be added to the production of the public good. Thus, in order to get people to participate in the revolution some reward must be offered to those who choose to participate which will not be shared with those who remain inactive. If Che would be offered a reward to compensate for the risk involved in participating, then the scale might be tipped in favour of revolutionary action. Alternatively, the same result could be achieved if he would be punished for remaining inactive.

This is represented in the formal model by adding a constant that represents the private benefit received by participating to equation (3.8).

$$U_{\text{participate}} = V(d) - C(d) + b \tag{3.9}$$

Although b might be too small to motivate revolutionary action if d is small, it might be large enough to compensate for the expected cost of participating as d increases. Once again, think of the Orange revolution in Ukraine in 2004 where rock concerts and soup kitchens were provided as private benefits for the participants. A bowl of soup was probably not a large enough benefit to persuade someone to single-handedly challenge the regime. However, the very same bowl of soup might have compensated for the relatively small risk of getting punished when being part of a large crowd.

Figure 3.6 shows a situation where it is actually rational to participate. In this case, if Che expects that more than $k-1$ other people will participate, then he will prefer participating to remaining inactive. The same will hold for everyone else. Furthermore, since nobody would gain by unilaterally switching from participating to remaining inactive if everyone else participates, the outcome where everyone participates is a Nash equilibrium. This does not mean that all-participate is the guaranteed outcome of the game. Consider Che again. If he believes that less than $k-1$ others would participate, then he will prefer to remain inactive. Since nobody would switch from remaining inactive to participating if everyone else remains inactive, the outcome where all remain

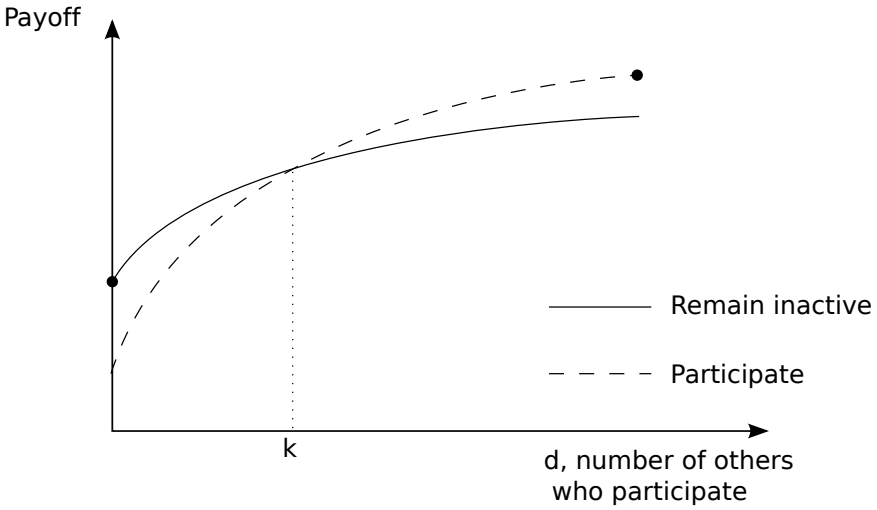

Figure 3.6: Payoffs with diminishing risk and private benefits, a *n*-player assurance game.

inactive is still a Nash equilibrium. That the game has multiple pure strategy equilibria is represented by the two black dots in figure 3.6. This game is sometimes called an assurance game or a coordination game, since the outcome of the game depends on the players' ability to coordinate.

3.4.1 Creating incentives

The payoff's tendency to converge can in conjunction with some selective incentives create a situation where a successful revolution is the outcome of rational action. Let us postpone, to the next section, the question about how coordination is achieved and first focus on the selective incentives and investigate under which circumstances they can be expected to exist. Although it is possible for us to just assume the existence of such incentives (making them exogenous to the model), an explanation where the incentives are a product of rational action (so that they become endogenous to the model) would be preferable.

Morris Silver [1974] argues that psychological benefits enter the potential revolutionary's payoff function in the sense that a feeling of duty, accomplishment, or belonging can motivate action. Although such psychological incentives can help transform the free-rider problem into a coordination problem, it seems difficult to show how a rational revolutionary in our model should go about to produce or internalize them. In other words, it seem as Silver has to take these incentives to be exogenous to the model. We will return to psychological incentives when we discuss the ideology theory of oppression in the

next chapter. There we will argue that the ideology theory can offer an explanation of how, e.g., a sense of duty can transform a prisoner's dilemma game into an assurance game.

Material incentives, on the other hand, seem to offer a way for the potential revolutionaries to manipulate their situation. Hardin [1995, p. 52] mentions punishment as one way of creating material incentives. Just as the regime is able to deter the oppressed from revolting by threatening to punish *if* they participate, so can the oppressed prevent their comrades from remaining inactive by threatening to punish them *unless* they participate. The result of successful threats would be similar to the situation depicted in figure 3.6.[1] The use of threats does, however, give rise to another problem: are the threats credible?

Unfortunately, the decision to punish comes after the decision to participate. Since punishing is usually costly, there seems to be little reason for an individual rational revolutionary to punish inactivity after the revolutionary outcome has been decided. The situation thus resembles the prisoner's dilemma situation since refraining from punishing dominates the choice of punishing. Thus, if everyone believes that the others are rational, they will not be swayed by the threats since they do not expect them to be carried out. The payoff function will remain unaffected.

It could perhaps be objected that it has been shown that people are actually willing to pay large costs to punish free-riders in public goods problems.[2] The problem is, however, how to make sense of this behaviour within the rational choice framework. We could do this by assuming the existence of retaliatory desires just as Silver assumed the existence of psychological incentives, or we could try to show how such behaviour can arise within the rational choice framework.

There are at least two ways of accounting for selective incentives within the rational choice framework: with the help of the existence of stable groups, or with a cleverly formulated agenda. Let us start with the stable group solution. Although punishing is strictly dominated in the one-shot game, it may become rational to carry out the threat if this interaction is embedded within a larger context. For example, since Fidel and Raúl are brothers they expect to run into each other at future family gatherings. They will do so whether the revolution succeeds or fails. Thus, if Fidel's reputation in future interactions depends on whether he carries out his threat, then he may have more to gain in the long run from punishing Raúl although it is costly in the short run. This line of reasoning is not limited to repeated interactions between the same two indi-

[1]The only difference would be that instead of shifting the curve representing the payoff for participation upwards, the curve representing the payoff for remaining inactive would be shifted downwards.

[2]See, e.g., Ernst Fehr and Simon Gächter [2000].

viduals. Assume, instead, that Che, Fidel and Raúl belong to the same political party. Then, if Fidel does not punish Raúl today, he may gain the reputation of being all bark and no bite. Thus, his future threats will be considered empty and rendered ineffective. On the other hand, if he carries out his threat, he may gain the reputation of being someone whose threats are to be taken seriously. Thus, his decision to punish Raúl today may affect his ability to threaten, e.g., Che tomorrow.

Therefore, if the decision to realise a threat is embedded in a larger context that implies repeated interactions, it can be rational to carry out a threat today in order to deter future deviations in the repeated setting even if doing so is dominated in the one-shot game. Being affected by such considerations is sometimes referred to as being in the shadow of the future.

Formally, we can transform Che and Fidel's dilemma from figure 3.2 to an assurance game by assuming that there is a probability, p, that the game will be played again. In these repeated games, we can assume that there are two salient strategies: Grim Trigger (GT) and Always remain Inactive (AI). The two strategies are specified as follows:

GT Participate the first time the game is played, then continue to participate unless the opponent has remained inactive. If the opponent has remained inactive once, then remain inactive in all future rounds.

AI Always be inactive.

In order to show that it can be rational to punish, i.e., play GT, we have to show that the payoff of playing GT is higher than that of playing AI. Given the payoff of the one-shot game we can write the payoffs for playing GT and AI as follows:

$$u(GT,GT) = 3+3p+3p^2+...+3p^n = 3\frac{1}{1-p}, \tag{3.10}$$

$$u(GT,AI) = 1+2p+2p^2+...+2p^n = 1+2\frac{p}{1-p}, \tag{3.11}$$

$$u(AI,GT) = 4+2p+2p^2+...+2p^n = 4+2\frac{p}{1-p}, \tag{3.12}$$

$$u(AI,AI) = 2+2p+2p^2+...+2p^n = 2\frac{1}{1-p}. \tag{3.13}$$

It is easy to verify that $u(AI,AI) > u(GT,AI)$ for all p and $u(GT,GT) > u(AI,GT)$ if $p > 0.5$. Thus, Che will always prefer to play AI if he believes that Fidel will play AI. If, on the other hand, he believes that Fidel will play GT *and* that there is a significant chance ($p > 0.5$) that the game will be repeated, then Che will prefer to play GT. In other words, as long as the probability

that the one-shot prisoner's dilemma will be repeated is higher than 0.5, the repeated game becomes an assurance game.[1]

Some contexts where the *shadow of the future* enters the picture are among villagers, members of trade unions, or circles of friends. The members of the groups can threaten those who do not participate with, e.g., corporal punishment, economic sanctions, or social sanctions such as being excluded from the next social event. Once a stable coalition has been formed, the revolutionary situation can be formulated as an assurance game where each member prefers to participate if she expects that the other members will participate.

Theda Skocpol's [1979] classic study of peasants in revolutions is a good account of the role groups play in achieving revolutionary success. Furthermore, Michael Taylor's [1988] description of the role played by political entrepreneurs in strengthening the peasant community in China indicates that revolutionaries have been aware of how important strong group ties are to revolutionary success.

A second suggested solution to the credibility problem comes in the form of an attempt to show that it is possible to transform the free-rider problem without having to assume repeated games and stable coalitions. Let us first make a distinction between different norm-levels. Given a game, $G =< N, S, U >$, a first-order norm says that the strategy, $s_i \in S_i$, ought to be played by player, $i \in N$. In Che and Fidel's dilemma (figure 3.2) the first-order norm might be that they ought to participate. A second-order norm is a norm telling the players how they ought to react to first-order norm violation and compliance. For example, that they ought to punish first-order norm violations and reward compliance. It is also possible to imagine higher-order norms that specify how a player ought to react to second-order, or even higher, norm violations. Higher-order norms are usually found in societies where there are strong norms of avenging wrongs. In such societies, it is not uncommon to also find a strong norm of punishing those who do not exact vengeance. The system of norms concerning revenge ascribed to medieval Iceland are sometimes used to illustrate higher-order norms.[2]

Jiborn [1999] observes that when first-order, second-order, and higher-order norms are offered one at a time, they create a free-rider problem. There will, after all, always be someone at the end of the chain who lacks an incentive to carry out the punishment. Since everyone knows this, each threat in the chain will lack credibility. If, on the other hand, the norms were offered simultaneously, as a package, then it is possible to overcome the free-rider problem.

[1]See, e.g., Bryan Skyrms [2001] and Jack Goldstone [1994, p. 143]. The latter applies the reasoning explicitly to revolutions. See also Robert Axelrod [1984] on the success of the famous *Tit-for-Tat* strategy in repeated prisoner dilemma games.

[2]For more examples and further discussion see, e.g., Elster [1990].

Jiborn [1999, p. 170] asks us to imagine the following two norms:

I. participate, and

II. punish each person who does not comply with norm I and II.

Let us name the strategy where a player complies with both norms *cooperate*, and the strategy where a player violates at least one of the norms *defect*. If very few choose to cooperate then a co-operator would have to punish many transgressions, which would be very costly. Thus an agent would prefer to defect when many others defect. On the other hand, if everyone else cooperates it will be costly to defect since everyone else stands ready to punish the single defector. Therefore a rational agent prefers to cooperate when many others are cooperating. Jiborn concludes that if the choice the agent faces is between complying and not complying with the package of norms, then the situation will resemble an n-player assurance game rather than an n-player prisoner's dilemma.

3.4.2 Coordinating the Stag Hunt

There are, in other words, reasons to believe that premise (b) of the rational choice argument at the end of section 3.3 is false, and that the revolutionary situation should be modelled as an assurance game rather than a prisoner's dilemma. This will at least be the case in oppressive societies where there exist stable groups and coalitions (assuming that some such group are also necessary for formulating a complex agenda). Thus, the proponent of the gunman theory of oppression can identify the existence of stable groups as a condition that, if fulfilled, increases the probability that a revolution will occur.[1]

However, unlike the n-player prisoner's dilemma game, where participation is strictly dominated by remaining inactive, we cannot predict how rational revolutionaries will act in a coordination game. Even if the rationality of the revolutionaries and the structure of the game is common knowledge, the revolutionaries cannot, without further information, deduce how the others will act. Each revolutionary's decision will depend on her beliefs about what the other revolutionaries' will do.

We can illustrate the underlying problem with the help of Che and Fidel's annual hunting trip. Each fall they go to the mountains where there are deer and hares to hunt. Since deer are rare and easily scared, the two friends need to cooperate to catch one. There are, on the other hand, plenty of hares which can be caught single-handedly. If both refrain from hunting hares, then there is

[1]However, see Nicolas Olsson-Yaouzis [2010] for some possible objections against the two solutions offered in the previous section.

a good chance that they will be able to catch a deer, and if both decide to hunt hares then both will return home with one hare each. If, however, one of them decides to hunt deer while the other hunts hare, the deer hunter will go home empty-handed whereas the hare hunter, who will have the hunting grounds for himself, will return home with two hares. Finally, returning from the hunting trip with a deer is strongly preferred to returning with a hare, which in turn is preferred to returning empty-handed.[1] This results in the payoff structure described by figure 3.7,[2] a two-player version of the n-player assurance game

		Fidel: Deer	Fidel: Hare
Che	Deer	9,9	0,8
	Hare	8,0	7,7

Figure 3.7: Che and Fidel's hunting trip.

described in figure 3.6. It is easy to verify that the strategy pair (Deer, Deer) and (Hare, Hare) are the only pure strategy equilibria of the game. The question becomes: When do Che and Fidel return with a deer and when do they return with hares?

One way of determining when they will return with a deer is to add the possibility of playing a mixed strategy. A mixed strategy is a strategy that involves randomisation over all or some of the available pure strategies.[3] For example, one mixed strategy would be to chase deer with the probability $\frac{1}{6}$ and hare with the probability $\frac{5}{6}$. Once we have introduced mixed strategies, we can calculate the mixed strategy Nash equilibria of the game: the set of mixed strategies no player would be better off unilaterally switching from.[4] Given the particular payoffs of the game described in figure 3.7 the mixed strategy

[1]The game has its origin in Rousseau's *A discourse on inequality* where he writes that "If it was a matter of hunting deer, everyone well realises 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." See Brian Skyrms [2001] for more details.

[2]The big payoff-difference between playing deer and playing hare when one's opponent plays hare, represents the risk of returning empty-handed. It is, in other words, much worse to switch from one hare to nothing, than to switch from a deer to a hare.

[3]Formally, a *mixed strategy* for player $i \in N$ is a probability distribution over i's set $S_i = \{1,2,...,m_i\}$ of pure strategies. Player i's mixed strategy, x_i, can be represented by a vector in m_i-dimensional Euclidian space, $\mathbb{R}^{m_i}$, where its hth coordinate $x_{i,h} \in \mathbb{R}$ represents the probability that the player i's hth pure strategy will be played. For example, one of Che's mixed strategies in the hunting trip is $x_{\text{Che}} = (\frac{1}{6}, \frac{5}{6})$.

[4]Formally, a mixed strategy profile $x = (x_i, x_{-i})$ is a Nash equilibrium if, and only

equilibrium strategy is to hunt deer with the probability $\frac{7}{8}$ and hunt hare with the probability $\frac{1}{8}$. It is easy to verify that if both play this strategy, then neither will be better off by unilaterally switching to another strategy. The expected payoff for playing the mixed strategy equilibrium will be $\frac{63}{8} = 7.875$ each.

If this was Che and Fidel's first hunting trip, or if they knew very little of each other, then the best response might be to play the equilibrium mixed strategy. This would allow them to return with a deer in about three out of four trips. A risk averse Che might, however, prefer to always hunt hare in order to ensure that he does not return empty-handed. Che may, for example, have a large starving family waiting for him to return with something to eat. For this reason the (Hare, Hare) outcome is called the *risk-dominant* outcome, whereas the (Deer, Deer) outcome is said to be the *payoff dominant* outcome.

Although playing the mixed Nash equilibrium might be a good idea in a game with multiple equilibria where no means of coordination exists, there are reasons to believe that in real world there are ways of coordinating. Consider, for example, Schelling's [1980] two parachuters who have been separated after a jump. Both have access to a map of the area, and they know that the other has access to the same map. There are a number of different items marked on the map: a number of farmhouses, some roads that intersect at a number of different places, and finally a river that can be crossed at *one* bridge. Although there are a near infinite number of equilibria in the game between the two jumpers, it seems more or less obvious where they will meet: the bridge. The uniqueness of the bridge makes it salient enough to cause the jumpers' expectations to converge on it. In Schelling's terminology the bridge becomes a *focal point* for the jumpers' expectations and as such it helps them to coordinate their actions.

If there was *something* that made Che and Fidel view hunting deer as the salient alternative, while at the same time making it clear that the other perceive the situation similarly, then deer hunting would become the focal point. A focal point can be man-made, e.g., when department stores provide a "lost and found" where customers who have become separated can meet, or they can be provided by nature, e.g., a clearing in a forest. All that is needed is that something makes an outcome, or a set of actions, salient enough for the players' expectations to converge on it.

In the deer hunting case, one seemingly obvious way of making deer hunting a focal point would be for Che to proclaim that he intends to hunt deer. If Fidel is able to reply, then it seems as his reply would ensure that deer hunting becomes the focal point. If we consider that both Che and Fidel prefer

if,

$$\forall i \in N, x_i \in X_i, x_i \neq x_i^* : u_i(x_i^*, x_{-i}^*) \geq u_i(x_i, x_{-i}^*).$$

the payoff-dominant outcome to the risk-dominant, then there seems to be no reason why either should play it safe and hunt hare once they have stated that they intend to hunt deer. Thus, communication, rationality, and the structure of the game being common knowledge seems to ensure that the payoff-dominant outcome will be played.

Robert Aumann [2000], however, has shown that if the structure of the stag hunt game and the players' rationality is common knowledge, then communication conveys no information about what the players will do. The reason is that no matter what Che has decided to do he will *say* that he will hunt deer. If Che has decided to hunt deer, then he would be better off if Fidel hunts deer; if he has decided to hunt hare, then he would also be better off if Fidel hunts deer since this would give him a payoff of 8 instead of 7. Thus there is no reason for him to ever say that he will hunt hare. Since their rationality and the structure of the game is common knowledge, both realise that the agreement means nothing and that communication is meaningless.

This theoretical result is difficult to reconcile with both intuitions and experimental results. For example, in one experimental study on the effect of communication on the outcome of Stag Hunt games Russell Cooper et al. [1992] find that without communication, coordination on the payoff dominant outcome is 0%, with one-way communication 53%, and 91% with two-way communication.

Tore Ellingsen and Robert Östling [2010] have shown that it is possible to reconcile the empirical findings with the rational choice theory if the common knowledge assumption is replaced by a *level-κ model of strategic thinking* where the players are allowed to doubt whether the other players understand the game.

The intuition behind the level-κ model of strategic thinking is that although players are not as sophisticated as described by the common knowledge assumption, they have a certain degree of sophistication. That is, although the players do not know that everyone else knows, that everyone else knows, that ...; the players have some beliefs about the other's beliefs. To be more precise, Ellingsen and Östling's model assumes hat the players have beliefs about whether the other players understand the game. For example, Che can believe that i) his opponent does not understand the game, or ii) that his opponent believes that he does not understand the game, or iii) that his opponent believes that Che believes that his opponent does not understand the game, or ...; where each belief corresponds to an increasing degree of sophistication.

Ellingsen and Östling show that if we add the assumption that the players have a weak lexicographical preference for sending truthful messages, then

communication can give the players information about the intended play.[1] The assumption that the players have a lexicographical preference for sending truthful messages entails that the players will always choose the message that maximises their expected payoff whether or not it is true; however, if a true and a false message would both maximise their expected payoff, then they will choose to send the truthful message. Although Ellingsen and Östling admit that there is considerable experimental evidence to indicate that real people have a stronger preference for sending truthful messages, they motivate the assumption of a lexicographical preference by pointing out that it will simplify the model. They also point out that if a stronger preference is assumed, then the results will remain the same. For our purposes the assumption of a lexicographical preference will suffice since it allows us to show that communication can be efficacious even under fairly weak assumptions.

The basic idea of the model is that there are different types of players and that some can be more sophisticated than others. A player of sophistication level $\kappa \geq 1$ is assumed to believe that her opponents are players of level $\kappa - 1$. For example, if Che is a level-2 player he will believe that Fidel is a level-1 player, and consequently, Che will believe that Fidel believes that Che is a level-0 player.

Furthermore, a player of the lowest level of sophistication, level-0, is assumed to uniformly randomise between the available alternatives. A level-0 player can be interpreted as a player who does not understand the game and therefore chooses an arbitrary strategy. In her view, all strategies have the same expected payoff. It is also possible to interpret the level-0 players as only existing in the minds of the more sophisticated players. On this interpretation, a player's belief that her opponent is a level-0 player can be seen as representing her fear that her opponent does not understand the game.

In order to see how communication can be efficacious in a world populated by level-κ agents, let us begin with the level-0 players and study their behaviour in a two-player stag hunt game. As we mentioned above, a level-0 player assigns the same expected payoff to all strategies. She will, therefore, randomise uniformly between playing deer and playing hare. Furthermore, since she does not believe that her messages will have any effect on the outcome of the game and she has a weak lexicographical preference for sending true signals, she will send a truthful message.

A level-1 player believes that her opponent is a level-0 player and does, therefore, believe that it will be equiprobable that her opponent plays hare or deer. If a level-1 player does not receive a signal from her opponent, then she will choose the strategy that gives her the highest expected payoff. Given the

[1] See Ellingsen and Östling [2010, p. 1698].

payoffs of the game described in figure 3.7 and her beliefs about her opponent, her expected payoff of playing hare ($0.5 \times 8 + 0.5 \times 7 = 7.5$) will be greater than her expected payoff of playing deer ($0.5 \times 9 + 0.5 \times 0 = 4.5$). She will, therefore, hunt hare. If there is one-way communication, and the level-1 player is the sender, then she will truthfully send the message that she will hunt hare.[1] If she is on the receiving end, and she is being sent a deer-hunting message she will adapt to, what she believes to be, the level-0 player's decision to hunt deer.

If two-way communication is allowed then the level-1 player will once again adapt to whatever signal she has been sent. Since she has a lexicographical preference for sending a truthful message, and she knows that she will adapt to (what she believes to be) a uniformly randomised message, she will send both messages with equal probability.

A level-2 player believes that his opponent is a level-1 player and will thus play hare in the absence of communication. If the level-2 player is unable to send a message and receives a message, she will adapt to whatever signal she receives. However, since level-1 players always adapt to the received messages, a level-2 player who is allowed to send a message will always send a deer-hunting signal and then hunt deer. For the same reason, if two-way communication is allowed, a level-2 player will send a deer hunting signal and then hunt deer no matter what signal she receives. Finally, it can be shown that all level-2_+ players behave like the level-2 player, implying that there will be coordination on the payoff-dominant equilibrium whenever two level-2_+ players that are allowed to communicate meet.[2]

Giving up the common knowledge assumptions should probably be seen as a feature rather than a flaw. This allows us to retain both the assumption that the revolutionaries are fully rational, and that communication is efficacious, while at the same time getting rid of a very controversial assumption.

We can now return to the question of when Che and Fidel will return with a deer and when they will return with hares. If they are unable to communicate, and they are risk averse (as described above) then they will both opt for the risk-dominant outcome and hunt hare. However, if both are sophisticated enough, level-2_+, then it is sufficient that one of them is allowed to send a message in order for them to coordinate on deer hunting. If two-way communication is allowed, then it is enough that one of them is a level-2+ player (and the other is not a level-0 player) in order for them to coordinate on the

[1] It can, perhaps, be questioned why a player who does not believe that she can affect an opponent who she does not believe understands the game, will bother to even send a message in the first place. In order to avoid this objection, let us assume that people have a weak preference for sending signals when they are allowed to.

[2] Ellingsen and Östling [2010, p. 1702].

payoff-dominant equilibrium.

Now, let us assume that three hunters are needed to bag a deer and that Che and Fidel decides to bring Raúl along. Assume first that multilateral communication is allowed. A level-1 player will hunt deer if all the received messages are deer-hunting signals and hunt hare if he receives at least one hare-hunting signal. Since a level-1 player believes that his opponent will uniformly randomise, he believes that the probability that both opponents will send him deer-hunting signals is $1/4$. Consequently, he will send a deer-hunting signal with the probability $1/4$ and a hare-hunting signal with the probability $3/4$.

A level-2_+ will play deer if all received messages are deer-hunting signals. Since a level-2_+ player understands that level-1 players who receive only deer-hunting signals will hunt deer, he will, in contrast to a level-1 player, always send a deer-hunting signal. Thus, if all players are level-2_+, then multilateral communication will guarantee that a deer is bagged no matter how many players are involved in the hunt.

Assume that Che, Fidel, and Raúl are level-2 player who would normally send a deer-hunting message, and assume that Fulgencio prevents Raúl's signal from reaching Che and Fidel. This means that everyone will receive two deer-hunting signals. In a way, this situation resembles a stag hunt with one-way communication. There is, however, one important difference between the two- and three-player scenarios. In the two-player scenario Che, a level-2 player, sends a deer-hunting signal. Che then hunts deer even if his opponent remains silent since he believes that his silent opponent, Raúl, believes that Che is a level-0 player who have chosen to hunt deer.

In the three-player scenario, however, Che has to worry about Fidel's as well as Raúl's beliefs. Since Che believes that Raúl believes that Che and Fidel are level-0 players, he still believes that Raúl will hunt deer. He will, however, have doubts about what Fidel will do. After all, he believes that Fidel believes that Che and Raúl are level-0 players. This means that Che believes that Fidel believes that Che will hunt deer, but it also means that Che believes that Fidel has no idea about what Raúl will do. Given the payoff structure of game 3.7, Che believes that Fidel will hunt hare. Consequently, Che will hunt hare. In other words, if they are prevented from communicating Che, Fidel and Raúl's hunting trip may easily be ruined.

If we return to the representation of the revolution as an n-player assurance game, we can point out some additional implications on this model concerning communication. Remember that we assumed that the revolution would succeed if at least k revolutionaries participate. When $n = k$ then the more signals Fulgencio can disrupt, the less likely a revolution will become. To see this, think of a level-1 player who has received revolutionary signals from everyone who can communicate, but has not heard anything from δ other revolution-

aries. The level-1 player believes that the probability that all of the δ silent revolutionaries will participate is $(\frac{1}{2})^{\delta}$. If $\delta = 1$ then there is a one-in-two chance that the silent revolutionary will participate. If the value of the successful revolution is high and the risk of getting hurt relatively small, then it might be rational to participate given these odds. As δ increases, however, the probability that $k-1$ others will participate in the revolution will rapidly decrease.

Thus, the more communication that is disrupted, the less likely will it be that a level-1 player believes that the revolution will succeed. The same is true for the case when $k < n$ since the probability that at least $k-1$ others will participate when $k < n$ also decreases in terms of δ. Therefore, if Fulgencio wants to prevent a revolution he should see to it that as much communication as possible is disrupted. The bottom line is that if the common knowledge assumption is replaced by the level-κ model of strategic thinking, then it is possible to make sense of the role of communication in revolutionary situations.

3.5 Explanation and false assumptions

One objection that could be raised against the gunman theory of oppression is that it rests on false assumptions. Everyone is not self-interested, egoistic, and behave as described by the level-κ model of strategic thinking. If the model is supposed to provide us with an *explanation* of persistent oppressive social orders, we will be forced to include false claims in our explanans. Thus, if we accept the DN-model of explanation, a false claim in the explanans will prevent us from deducing the explanandum, and, according to Hempel, from providing an explanation of the phenomenon we are interested in. A proponent of the ideology theory of oppression might therefore be tempted to argue that this gives us conclusive reasons to abandon the gunman theory of oppression.

Since economic models often include false assumptions, this problem has long haunted economists. One of the most influential ways of handling the problem among economists was provided by Milton Friedman [1994] who suggested that economic models should not be judged by how well the assumptions describe reality. Rather, economic models should primarily be judged by their predictive power. On this account, it would not matter whether the gunman theory of oppression assumed the existence of ghosts and goblins as long as they would help us predict the rise and fall of oppressive social orders.

However, the problem is that the predictive success of economic and rational choice models is somewhat underwhelming. Economists did, for example, fail to foresee both the subprime mortgage crash and the Euro crisis, and social scientists failed to predict the Arab spring and the collapse of the Soviet Union. Furthermore, as Alexander Rosenberg [2008] argues, if economic and rational

choice models were successful in generating predictions, then the best explanation for this seems to be that the models are in fact true or close to being true. This explanation would, however, bring us back to a realistic interpretation of the economic and rational choice models.

Also remember that one of the reasons that analytical Marxists prefer the gunman theory to the ideology theory of oppression is that it allows them to avoid becoming committed to the existence of ontological extravagant entities. Although they would probably agree that rational choice theory allows them to make better predictions, its predictive power is not the main reason why they opt for the gunman theory.

Another way of accounting for the explanatory power of models using false assumptions is provided by Robert Sugden [2000], who argues that we should view models as describing counterfactual worlds. The model provided by the gunman theory describes a counterfactual world where everyone are self-interested, egoistic, and behave as described by the level-κ model of strategic thinking. Thus, it can be said that our model provides an explanation of persistent oppressive social orders in this counterfactual world.

Whether the model has any explanatory power in the actual world depends, according to Sugden, on how *credible* the counterfactual world is. Without going into the details we can say that a model world, v, is more credible than another world, v', if, and only if, v is closer to the actual world than v' is.[1] On this account, a model that assumes the existence of ghosts and goblins would (all other things being equal) have a lower explanatory power than a model that does not make this assumption since the latter model world is more credible than the former.

It should, however, be pointed out that a high degree of credibility does not trump all other explanatory values. If it did, then the model that described the actual world in its every detail and left nothing out would be the best model. There are, however, reasons to prefer less credible model worlds. A less credible model world might, for example, be preferable since it provides us with, or calls our attention to, more relevant information concerning the phenomena we are interested in. For example, Hal Varian and Alan Gibbard [1978] argue that economic models should be viewed as caricatures. Just as a caricature portrait draws our attention to certain facial features at the expense of verisimilitude, so do economic and social scientific models make salient certain relevant features of the world at the expense of credibility.

For example, since the actual world is inhabited by all sorts of people, not only self-interested egoists following the level-κ model of strategic thinking, our model provides a simplification of the actual world. However, if *most*

[1] Where closeness has to be specified in some appropriate way.

people are egoistic, or at least egoistic enough, then this simplification will not detract from the explanatory value of the model since it provides us with a caricature where the *relevant* social features have been made salient.

In the end, it boils down to what we value in an explanation. It is, of course, possible to treat true assumptions (or maximal credibility) as an absolute value. The problem is that this will disqualify not only rational choice explanations, but also explanations employing, e.g., Newtonian mechanics. For now, let us note that this is a high price to pay. We will return to the question of what characterises a good explanation in chapter 5.

A more fruitful objection to the gunman theory of oppression is that although it does a good job at explaining tyranny, such as the case of North Korea or East Germany, it does a worse job at explaining the other cases of persistent oppressive orders. We will turn to this issue in the next chapter.

3.6 Summary

In this chapter we have examined Tocqueville's argument for why a few guards at Sing-Sing could hold a large number prisoners at bay. We also showed how it can be successfully applied on some persistent oppressive social orders as well.

We have seen that it seems as the gunman theory of oppression does a very good job at explaining persistent tyrannies. It can offer an explanation as to why the North Koreans do not stage a revolution against their oppressors that successfully accounts for many of the interesting features of this oppressive social order. Furthermore, it allows for the possibility that the North Koreans understand that they are oppressed and that there exists an alternative less oppressive social order. The theory also accounts for the fact that the regime goes to great lengths to prevent people from freely expressing their opinions. Finally, it explained the fall of other oppressive social orders in terms of communication and coordination.

The version of the gunman theory of oppression that we have discussed in this chapter should be seen as one version of a rational choice explanation of persistent oppression. It is, after all, possible to use rational choice theory to explain continued oppression where the oppressed are not literally held off at gunpoint. What the different versions have in common is that they all attempt to explain the persistence of oppressive social order by pointing to the correct beliefs and 'autonomous' desires of the oppressed. More precisely, they claim that all persistent oppression can be explained by the following belief-desire pair: the oppressed believe (correctly) that it is individually costly to attempt to exchange the oppressive social order for a less oppressive social order, and the oppressed have a desire to maximise their own welfare.

In the next chapter we will investigate whether there is a version of the gunman theory of oppression that can be used to explain the other cases of oppression that were introduced in section 2.5.

4. The explanatory importance of ideology

4.1 Introduction

> In the social production of their life, men enter into definite relations that are indispensable and independent of their will, relations of production that correspond to a definite stage of development of their material productive forces. The sum total of these relations of production constitutes the economic structure of society, the real foundation, on which rises a legal and political superstructure and to which corresponds definite forms of social consciousness. The mode of production of material life conditions the social, political and intellectual life process in general. It is not the consciousness of men that determine their being but, on the contrary, their social being that determines their consciousness. [Marx, 2001, p. 7].

Karl Marx's famous base-superstructure remark has been analysed, deconstructed, and reconstructed. We will ignore both Marx's own remarks and (most of) his commentators.[1]

What is at stake is not whether Marx got the details right, but rather whether ideology of some kind is needed to give a full explanation of persistent oppression. After all, if the gunman theory of oppression can offer satisfactory explanations of all interesting aspects of oppression, then it is difficult to motivate why we should introduce yet another theory.

There are, however, plenty of cases of oppression for which the gunman theory is unable to offer straightforward explanations. The gunman theory of

[1] One of the reasons is that it would take a book of its own to figure out what Marx meant, and a couple of more to figure out how 'ideology' has been used during the 20th century. Michael Rosen [1996], for example, count to six different uses of 'ideology' in Marx's writings alone, and Terry Eagleton's [2007] excellent account of the development of the word shows that the use of 'ideology' during the French enlightenment has little in common with how it was used by, e.g., 20th century French post-modernists.

oppression is, after all, mainly used for explaining the persistence of, what Iris Marion Young [2005] calls, tyrannical social orders: regimes where a small number of rulers brutally oppress a majority. The oppression that Young [2005, p. 41] is interested in, however, "designates the disadvantage and injustice some people suffer not because a tyrannical power coerces them, but because of the everyday practices of a well-intentioned liberal society." That is, oppressive social orders where it is difficult to identify a (literal) gunman.

In this chapter we will provide an argument for the explanatory necessity of ideology along the lines of Durkheim's [1979] argument for the existence of social entities. He began by showing that it was impossible to explain the sharp increase in suicides between 1856 and 1878 in terms of individual and psychological facts. He argued that since social facts were part of the best explanation of the increase in suicides, social facts must exist. However, we will not follow Durkheim and conduct a rigorous empirical study. Instead, we will examine some suggested explanations that share the gunman theory of oppression's assumptions and show that they fail to provide satisfactory explanations of the persisting economic inequality in the US and the gender inequality in Sweden. Once this is shown, we will proceed to show that it is possible to account for the persistence of these social orders with the help of ideology.

More specifically, in section 4.2, we will focus on two attempts to explain why an unequal social order, of the kind described in section 2.5.2 under the heading of *The American dream*, persists. It will be argued that although the two explanations capture some of the aspects of this case, it fails to offer a straightforward account of other important aspects.

Its failure to do so provides us with a reason to investigate whether the ideology theory of oppression can offer a better explanation. In section 4.2.1, we will introduce the concept of ideology and show how it can help explain persistent oppression. In order to avoid disappointment, however, let us point out that neither necessary and sufficient conditions for ideology, nor an exhaustive list of all relevant types of ideology will be provided. For our purposes it is sufficient to show that ideology (in some form) is necessary to explain one of the cases that the gunman theory of oppression fails to explain. We will offer two examples of ideology: false beliefs and ideological norms. We will show how false beliefs can explain the persistence of the inequalities in USA as they were described in section 2.5.2. Ideological norms, on the other hand, will be used to explain the persistence of the gender wage gap in Sweden as it was described in 2.5.3.

Once we have indicated that ideology is needed to explain the persistence of oppressive social orders we will, in section 4.3, formulate the ideology theory of oppression. We will make a distinction between two versions of the ide-

ology theory of oppression: one that claims that ideological beliefs and norms explain persistent oppression, and another that adds the claim that ideological beliefs and norms exist because they *serve the function* of preserving the oppressive social order. Although we will postpone the defence of the functional claim to chapter 5, we will in this section take care of some preliminary objections. Section 4.4 concludes.

4.2 Oppression without gunmen

A complete theory of persistent oppressive social orders should be able to explain why the extremely unequal US social order described in section 2.5.2 persists. At first sight it might seem as the gunman theory is unable to even attempt an explanations of this phenomenon. After all, the US is, despite its flaws, a relatively well-working democracy that (at least as of late) does not shoot at protestors and demonstrators. Thus, it would be difficult to accept the gunman theory if it implied that the economic inequality in the US persist because the Americans are held off at gunpoint.

This objection would, however, rest on an uncharitable interpretation of the gunman theory of oppression. As we mentioned at the end of chapter 3, the gunman theory of oppression attempts to explain persistent oppressive social orders with the help of the assumption that the oppressed correctly believe that it is individually costly to attempt to exchange an oppressive social order for a less oppressive order, and that all oppressed have a desire to maximise their own well-being. Since the cost of participating does not have to involve tear gas, torture, or death, it might be possible to formulate a version of the gunman theory of oppression that does not literally involve gunmen.

One interpretation of the gunman theory can be attributed to Adam Przeworski [1980] who argues that even if there are no actual gunmen there are still individual costs associated with exchanging an oppressive social order for another. Przeworski attempts to explain why workers do not opt for socialism in Western democracies. He argues that even if 1) workers are interested only in satisfying their material interests, 2) workers correctly believe that socialism would better serve these interests than capitalism, and 3) workers correctly believe there is nothing preventing them from exchanging a capitalist order for a socialist order, workers would still not opt for socialism.

The reason, according to Przeworski, is that there is a transaction cost involved in exchanging one social order for another that must be paid by the generation of workers who experience the system shift. If the cost is large then this generation of workers will become worse off than under the capitalist system. So although every future generation of workers would be better off, the transaction cost will deprive the present generation of any incentive to ex-

change capitalism for socialism. Since this will be true of each generation of workers, there will never be a system shift.

There seems to be something to the claim that there are large transaction costs involved in exchanging one economic or political system for another. While relatively unbloody, the transitions of Russia, Poland, and East Germany from plan economies were far from painless. Likewise, although it is perhaps too soon to say, the North African countries will be in for some years of economic setbacks due to systemic change. Therefore, it is reasonable to assume that a transition to socialism would involve a similar transaction cost.

Applying Przeworski's argument to the persistent American inequality, the lack of social unrest will be explained by the fact that political reform would make the present generation of American poor even worse off than they are today. Since the transaction cost removes the incentive for revolution or reform, we should not expect an American social revolution anytime soon. Consequently, it is not surprising that the unequal social order persists.

What is mysterious about persisting oppression, however, is not only that the oppressed do not opt for the least oppressive social order, but also that they do not even struggle for some social order that is less oppressive than the present order. Przeworski might have offered an explanation of the former phenomenon by convincing us that the present generation of workers will be worse off by switching to the least oppressive social order, i.e., socialism. However, he will have some work to if he wants to convince us that all social reform would result in a worsening of the welfare of the present generation of workers.

As it stands, his argument presupposes either that there is only one alternative to capitalism or that all alternatives involve an equally high or higher transaction cost. Concerning the first option, it seems unlikely that the only alternative to the American model is socialism. It is, for example, difficult to see any real reason why the Canadian model would be unavailable for the Americans. Concerning the second alternative, it seems unlikely that the transaction cost of slightly increasing taxes for the richest 1% will be so high that the net result for the poorest will be that they were better off without the tax increase.

A second explanation of why oppression persists in democratic societies is provided by Torbjörn Tännsjö [2006]. He makes similar assumptions as Przeworski and argues that revolutions or reforms should not be expected even if:: workers are self-interested, and correctly believe that they can bring about a less oppressive social order and that no "gunmen" will try to stop them. The reason, according to Tännsjö, is that the nature of the globalized economy allows capitalists to move their capital freely across borders. This ability makes it possible for the capitalists to blackmail workers into accepting an oppressive social order. The argument can be illustrated with the help of the game de-

picted in figure 4.1 between the worker majority and the capitalist minority.[1]

		Workers: Unjust	Workers: Just
Capitalists	Unjust	8, 2	6, 1
	Just	0, 0	5, 5

Figure 4.1: Laissez-faire vs. egalitarianism game.

If both capitalists and workers opt for the unjust system, then capitalists will go on exploiting workers. If both opt for a just system, capitalists will be worse off and workers better off than in the unjust social order. Note that in this model there will be no loss of total productivity when switching economic system. Furthermore, note also that if the social order is changed, workers will be immediately made better off. Thus, we do not have to worry about Przeworski's transaction costs.

However, if workers attempt to unilaterally exchange the unjust system for a more just, capitalists will send their money to an offshore account on the Cayman Islands. We can assume that capitalists will lose some of their capital due to, e.g., temporary production losses, whereas the workers who become deprived of a necessary means of production will lose much more. Finally, although it is unlikely that capitalists will unilaterally opt for a just society, let us for the sake of completeness assume that the payoff would be zero to both workers and capitalists.

Applied to the mystery at hand we could say that if the American poor would start to talk about justice and equality, the American (super-)rich would threaten to escape with their capital. Since the poor realise that the rich have this option, they will refrain from demanding reforms and more equality. Therefore, the oppressive social order persists

However, Tännsjö's argument rests on the same empirical assumption as Przeworski's argument: either there is only one alternative or all alternatives have the same payoff consequences.[2] Once again, concerning the first option it is far from obvious that the only alternative to the American model is egalitarianism.

Concerning the second option consider the game in figure 4.2 where the

[1] Tännsjö uses a clash of wills game to illustrate his point. We will, however, use a game where the decision to remain unjust dominates the strategy of going just for the capitalists in order to make it clear that the capitalist's threat to move their capital abroad is credible.

[2] Or to be more precise, all alternatives have the same payoff consequence with respect to an ordinal ranking of the outcomes.

players choose between unjust and less unjust. If both capitalists and workers

		Workers	
		Unjust	Less unjust
Capitalists	Unjust	8, 2	6, 1
	Less unjust	0, 0	7, 4

Figure 4.2: Laissez faire vs. social democracy.

opt for the less unjust alternative, then although capitalists will be less well off than they are in the unjust social order they will still be much better off than the workers. However, and this is important, they will be better off in the less unjust social order than on the Cayman Islands. If capitalists are rational, then since they are better off in the less unjust social order than on the Cayman Islands, their threat to escape with the capital will not be credible. If this is a correct representation of the American situation, then Tännsjö's model fails to explain the persistence of the economic inequality. The mystery will, in other words, persist.

We must admit that both Tännsjö's and Przeworski's models can be used to explain some possible cases of persistent oppression without (literal) gunmen. However, if Tännsjö or Przeworski wants to make the bolder claim that the persistence of some actual persistent oppressive social orders can be explained by their models, then they have to provide us with good reasons to believe that the real world resembles their model worlds with respect to payoffs and transaction costs.

The problem is that there are reasons to doubt that their models accurately describe the case of, e.g., the persistent American inequality. After all, given that there are many non-socialist and non-egalitarian social orders with higher social mobility and equality than the US, it is hard to accept that there can be no other alternatives to the American model than socialism or egalitarianism.

Furthermore, recall the NYT study from section 2.5.2 where it was shown that many Americans falsely believe that their society has a high level of social mobility. Even if Tännsjö or Przeworski could provide evidence for the claim that their models accurately describe the dilemma facing the American poor, they would still have to account for this false belief held by the Americans. As we shall see, the ideology theory of oppression offers a relatively straightforward account of this phenomenon by explaining the persistence of oppressive social orders by citing ideological beliefs. This is something the gunman theory is unable to do since it assumes that the oppressed correctly believe that they are oppressed.

Tännsjö [2006, p. 429] attempts to account for ideology by arguing that both the persistence of oppression and ideological beliefs are caused by the in-

ability to solve collective dilemmas. According to him, this inability gives rise to ideological beliefs through a psychological mechanism that minimises cognitive dissonance. Therefore, when the members of the oppressed group realise that they cannot satisfy their desire for a more equal society they deal with their dissonance by forming the belief that their society is more equal than it actually is. Furthermore, according to Tännsjö, if the collective dilemma would be solved, the ideological beliefs would immediately disappear. On Tännsjö's account, ideology becomes causally irrelevant when it comes to explaining persisting oppression. We will postpone the discussion of this argument to section 4.3 after we have introduced the ideology theory of oppression.

4.2.1 Ideology

In a nutshell, the ideology theory of oppression can be said to deny the claim that the persistence of all oppressive social orders can be explained by assuming i) that the oppressed have correct beliefs about their situation and ii) that they desire to maximise their own interest-satisfaction. However, this does not mean that the ideology theory of oppression assumes that, as Eagleton [2007] puts it, the oppressed are cynics, that feel no unease about being oppressed, or are masochists, that enjoy it. Rather, the ideology theory of oppression rests on the assumption that "the majority of people have a sharp eye to their own rights and interests, and most people feel uncomfortable at the thought of belonging to a seriously unjust form of life."[1] The function of *ideology* is to keep the oppressed in the dark about their oppression. It does this by causing the oppressed to believe that, e.g., the injustices they are suffering are en route to being corrected, or that they are counterbalanced by greater benefits, or that they are inevitable, or that they are not really injustices at all, or that the injustices are so small that there is no need to bother with taking action.

The example of the persistent economic inequality in the US seems to be a paradigmatic case where oppression is upheld by ideological beliefs. As we saw in the previous section, the attempts to use the gunman theory of oppression to explain the persistent inequality ran into problems. Among other things, the gunman theory of oppression failed to account for the fact that many American citizens believe that their society has a higher degree of social mobility than it actually has. The explanation offered by the ideology theory of oppression is straightforward: there is no civil unrest among America's poor because they do not believe that they are oppressed.

In this case, the ideology theory of oppression accepts the gunman theory's assumption that the oppressed are motivated to maximise their payoff in terms of welfare. If a person is motivated to maximise her own welfare, then she will

[1] Eagleton [2007, p. 27].

be motivated to bring about social change only if she believes that there exists an alternative social order that is better for her own welfare than the present social order. The ideology theory of oppression goes on to point out that the false belief that there is a high degree of social mobility causes the citizens who desire a high degree of social mobility to believe that there is no better alternative, and thus removes their motivation to bring about social change. This explanation follows the standard Humean model where actions are explained by a belief-desire pair: the desire to maximise own welfare combined with the ideological belief cause the oppressed to abstain from revolutionary action.

The revolutionaries are, in other words, rational in the sense that they want to maximise their own welfare. They fail to do so, however, as they falsely believe that the current social order is welfare maximising. Assume, for example, that Che, the rational revolutionary, desires to live in a society where the children of the poor have the same chance to succeed as the children of the rich. Fulgencio, however, is content with the present system where children tend to inherit the economic and social status of their parents. Fulgencio can prevent Che from participating in revolutionary activity by convincing him that they live in a society where "each man and each woman [are] able to attain to the fullest stature of which they are innately capable and be recognised for what they are, regardless of the fortuitous circumstances of birth positions."[1] If Fulgencio can convince Che that they live in a just society, then there will be no need to point a gun at him or to prevent him from communicating with Fidel.

We can use the fact that the two theories, in this case, share the basic assumptions of rationality to construct a model that shows the effects of ideology on revolutionary situations. In order to do this, let us begin by making a distinction between interests and desires: something is said to be in the *interest* of a person if it would increase her welfare, and a person is said to have a *desire* (or preference) for something if she is motivated to get it. This distinction allows us to distinguish between games where the payoffs are given in terms of welfare and games where the payoffs are given in terms of desire or preference satisfaction. Although the first type of games gives us the outcomes that would make the players' lives go well, only the latter allows us to explain and predict their behaviour.

For example, suppose that the n-player stag-hunt game from figure 4.3 is

[1]The quote is taken from James Truslow Adams [1931, p. 404] who defines the American dream as the "dream of a land in which life should be better and richer and fuller for every man, with opportunity for each according to ability or achievement. [...] a dream of social order in which each man and each woman shall be able to attain to the fullest stature of which they are innately capable and be recognised for what they are, regardless of the fortuitous circumstances of birth or position."

offered as a description of a revolutionary situation.[1] A proponent of the ideol-

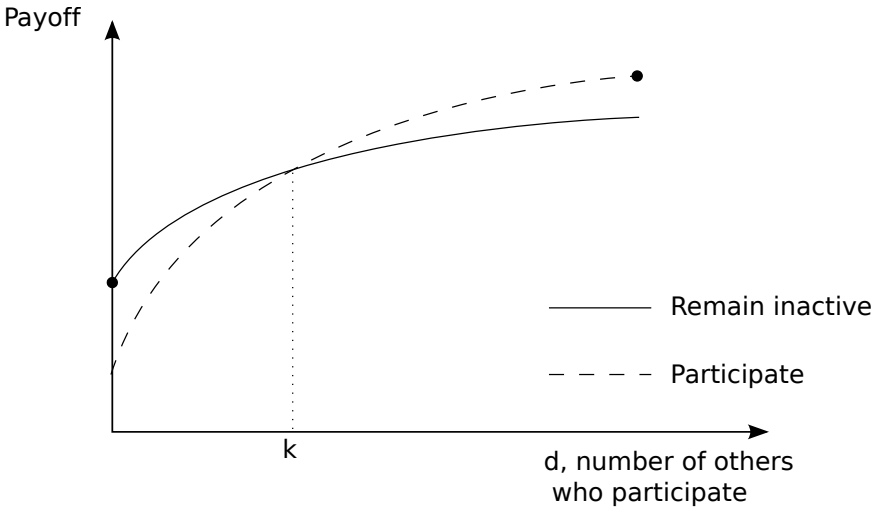

Figure 4.3: Payoffs represent interest-satisfaction, a *n*-player assurance game.

ogy theory of oppression can accept that the *n*-player stag-hunt correctly represents the revolutionary situation if the payoff function describe each player's level of *interest-satisfaction* as a function of the number of participants. Thus they can accept that it is in Che's *interest* to participate if at least $k-1$ other players participate, and against his interest if less than $k-1$ other players will participate. Furthermore, they can accept that the best thing (in terms of welfare) that could happen to Che is that everyone participates.

The problem is that interests cannot directly explain the revolutionaries' actions. For example, the facts that being in good health is in Fidel's interest and that smoking is bad for Fidel's health, does not prevent Fidel from desiring a cigar. The gunman theory of oppression connects desires with interests by assuming that the interests and desires of the oppressed coincide. Although the ideology theory of oppression denies this assumption, it agrees that desires are relevant for action explanations. Let us therefore represent ideology as a 'filter' that enters between the potential revolutionaries' interests and desires, so that ideological beliefs transform *interest-games* into *desire-games*.

$$I : G_{\text{interests}} =< N,S,U > \mapsto G_{\text{desires}} =< N,S,U' > \tag{4.1}$$

We can illustrate this with the help of Che and Fidel's hunting trip as shown in figure 4.4. Once again, Fulgencio desires to ruin the hunting trip. This time, however, he decides to use his newspapers and TV stations to influence Che. He orders the newspapers to report that there are plenty of hares in the

[1]Figure 4.3 is identical to figure 3.6.

mountains and the TV news to run shows about hunting teams who set out to hunt deer but returned disappointed and empty-handed. Although Fidel stays unaffected by the propaganda, Che's beliefs are affected. He now believes that it will be extremely difficult to catch a deer and easy to catch a bunch of hares. The new beliefs will cause him to strictly prefer hunting hare to hunting deer. Their hunting trip can now be represented by the (desire-)game shown in figure 4.5.

		Fidel: Deer	Fidel: Hare
Che	Deer	$4,4$	$0,2$
	Hare	$2,0$	$1,1$

Figure 4.4: Che's and Fidel's new hunting trip. Payoffs represent interest-satisfaction.

		Fidel: Deer	Fidel: Hare
Che	Deer	$1,4$	$0,2$
	Hare	$4,0$	$3,1$

Figure 4.5: Che's and Fidel's new hunting trip. Payoffs represent desire-satisfaction.

Note that although Fulgencio has failed to convince Fidel, he has, by successfully changing Che's beliefs, effectively ruined their trip. After all, if Fidel realises that Che has been deceived, then it will be rational for Fidel to hunt hare in order to avoid returning empty-handed.

Let us move on to the revolutionary situation described as a n-player stag-hunt game. A widespread belief, such as the American dream, that the social order is much better than it actually is transforms the n-player stag-hunt interest-game (by changing the public perception of the alternatives) into a desire-game with a unique equilibrium where nobody engages in revolutionary action. For example, a person who desires high intergenerational social mobility and believes that the present social order has an optimal degree of social mobility will consider a revolution to be meaningless at best, and catastrophic at worst. In her view, participating in a revolution will not only put her in harms way, it will also risk making society worse than it is. Since she believes that she has absolutely nothing to gain, remaining inactive will dominate the alternative no matter how many others are expected to participate. This game can be seen in figure 4.6.

Just as it was unnecessary for Fulgencio to deceive both Che and Fidel

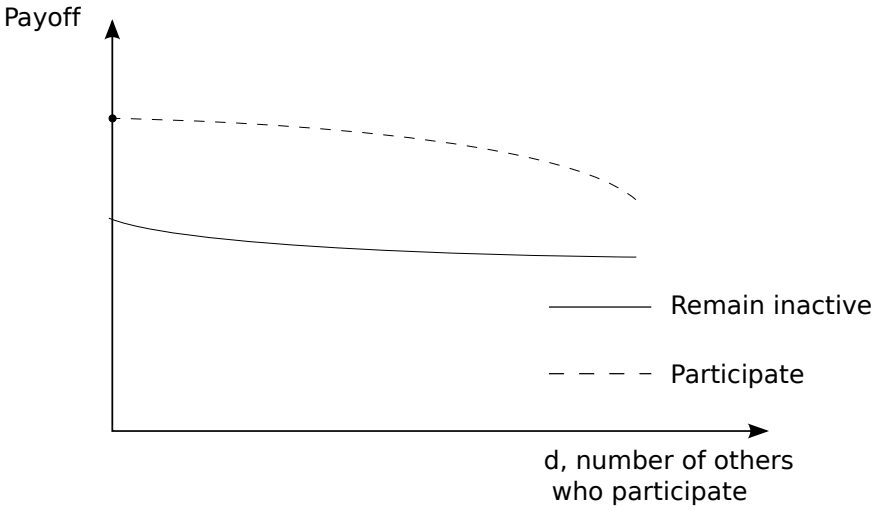

Figure 4.6: Payoffs represent desire-satisfaction, remaining inactive dominates participating.

in the two-player case, it is unnecessary for everyone to be deceived in the n-players case in order for the revolution to fail. Since (by assumption) k revolutionaries have to participate in order for the revolution to succeed, it is enough that at least $n-k+1$ potential revolutionaries are deceived for the revolution to fail. Since $n-(n-k+1)<k$, no coordination effort in the world will cause a rational revolutionary to participate. Furthermore, if we assume that the difficulties of coordinating a revolution increase as the number of potential revolutionaries decrease,[1] then each victim of ideology will lower the probability of a successful revolution. This suggests that it can be meaningful for an oppressive regime to spread ideology even if they will be unable to reach the magic number of $n-k+1$ victims. It also seems to explain why oppressive regimes spend resources both on propaganda and on security forces.

This means that it is possible to formulate the ideology theory of oppression so that it does not imply that the oppressive status quo is maintained because *everyone* are victims of ideological beliefs. It is sufficient that the members of a substantial part of the citizenry are victims of ideology. The rationality and the correct beliefs of the unaffected members of the oppressed majority will take care of the rest.

It should also be noted that we can use a similar framework to make sense of the psychological benefits that were introduced in chapter 3. There we suggested that revolutionary free-rider problems can be overcome if people re-

[1] One reason may be that when there are less potential revolutionaries to recruit, they will be more difficult to find.

ceived psychological benefits from engaging in revolutionary activities.[1] A strong sense of duty, accomplishment, or belonging could, for example, compensate an agent for the individual risk she would take by participating in a revolution. Within the framework presented in this section, psychological benefits would be represented as ideology-like beliefs that transform the perception of the revolutionary free-rider problem into, e.g., a *n*-player stag hunt. In other words, ideology-like beliefs can enter the picture, not only to prevent revolutionary activity, but also to facilitate it by making the revolutionaries willing to act against their immediate interests. For example, a revolutionary ideology can cause revolutionaries to make huge sacrifices by, for example, getting them to believe in an afterlife, aligning their interests with the interests of the collective, or making them idealise the post-revolutionary social order.[2]

Before moving on we should point out two things. First, although we have offered a sketch of how ideological beliefs can be used to explain persistent oppressive social orders, we have not said anything about where such beliefs come from. This is something we will address in section 4.3 when we discuss Tännsjö's claim that ideology is irrelevant with respect to explanations of persisting oppression.

Second, it should be mentioned that although ideology is often taken to imply *false* beliefs this does not necessarily mean that all ideology involve false beliefs.[3] It might, for example, be argued that Fulgencio's belief that "schools in residential areas with a lot of immigrants are usually worse than schools in areas inhabited mainly by ethnic Swedes" may be true and yet be an example of an ideological belief since it seems to imply that immigrants *qua* immigrants make the schools worse.[4] Another point worth making is that just because a belief is false does not mean that it is ideological. For example, that Fulgencio is mistaken about having four coins in his pocket does not imply that he is a victim of ideology.

However, since we are neither aiming at a definition of ideology nor a conclusive list of ideological beliefs we will not dwell on these points. For our purposes, it is enough to have shown i) that rational choice theory runs into problems when it comes to explaining the lack of civil unrest among America's poor, and ii) that the false belief about a high degree of intergenerational social mobility in the USA can be used to explain this lack of revolutionary activity.

[1] See also Silver [1974].

[2] Tännsjö [2007, p. 233] has, for example, suggested that the belief that an unequal society has a *lower* social mobility that it actually has can, by raising people's "class-consciousness," cause them to engage in revolutionary activities.

[3] Ideology is sometimes referred to as 'false consciousness.'

[4] See, e.g., Eagleton [2007, p. 15-6].

4.2.2 The gender wage gap

Although the explanation of persistent inequality in the USA in terms of ideology is intuitively appealing, it is difficult to immediately identify the corresponding false belief that upholds the gender wage gap in, e.g., Sweden.[1] One suggestion might be that women falsely believe that they earn as much as men. Most women, however, are painfully aware of the fact that they have a lower earning power than their male counterparts. A second suggestion might be that they falsely believe that women are paid less than men because being a woman is correlated with other relevant factors: a preference for a type of work that happens to pay less, a preference for more flexible working hours, or a tendency to yield in bargaining situations. The problem is that it seems to be true that much (although not all) of the gender wage gap can explained by these factors. Although it is possible that ideology can consist of true beliefs, we will argue instead that gender inequality is caused by the internalization of norms that prescribe how women and men ought to behave. Let us, however, begin by having a look at an explanation of the persistent gender inequalities in Sweden that uses similar assumptions as the gunman theory of oppression.

Bo Rothstein [2010] offers, what he calls, a causal mechanism approach to the persisting gender inequality in Sweden. Remember that we in section 2.5.3 saw that Browne [2002] explained the gender wage gap with the help of women's tendency to value flexible work hours and choice of education. Instead of explaining the gender wage gap by referring directly to life choices, Rothstein offers a micro-mechanism that shows that these choices are the outcome of rational self-interested behaviour.

The mystery, according to Rothstein, is why gender inequality persists in a country where almost everyone agrees that gender equality is something good and desirable. The situation becomes yet more mysterious if we consider that Sweden has introduced a number of progressive policies with the intention of increasing gender equality. For example, in order to remove the potential inequality caused by an unequal division of parental leave, Swedish fathers have been given the same right to parental leave as Swedish mothers. Yet gender inequality persists. According to Statistic Sweden, in 2010, mothers used 77% of the parental leave. When a child became sick, the mother stayed home 64% of the time. And in 2009, women earned on average 85% of what men earned.[2]

Rothstein's argument begins with the observation that men are on average

[1] Described in section 2.5.3.

[2] http://www.scb.se/Pages/ThematicAreaTableAndChart____327790.aspx and http://www.scb.se/Pages/ThematicAreaTableAndChart____327792.aspx accessed on March 6, 2012.

three years older than women when heterosexual couples are formed. Being three years older means that the man will have a somewhat stronger position on the labour market than the woman. Rothstein goes on to argue that if the couple is rational and desires to maximize household income, the stronger position on the labour market will translate into a stronger bargaining position if they become parents. When the woman loses the first initial negotiation about how to divide the responsibility of the household work, her position on the labour market will become even worse with respect to future negotiations. After each new bargaining situation the man will become better and better situated on the labour market, whereas the woman's position will become worse and worse. Thus, even if the initial difference between housework and paid work is small, it can in time become significant.

In support for his model Rothstein [2010, p. 13] presents answers to a questionnaire that shows that among (heterosexual) couples with children in Sweden, housework tends to be more unequal if the man is at least three years older than the woman. For example, among the women in these couples, 42% claimed that they spent much more time on housework than their spouse, 21% said that they contribute with one-quarter or less of the family's total income, and 19% agreed completely when asked if they feel that is their duty to do most of the household work. The corresponding number for women in couples where the man was less than three years older than the woman was 34%, 12%, and 12%.

In order to explain the gender wage gap it is sufficient for Rothstein to assume that employers are aware of the typical outcome of couples' negotiations. Given this assumption, employers will expect women to be home from work more often than men. Consequently, female employees will, in the eyes of employers, be worse investments than male employees, and therefore paid lower wages. Rothstein's explanation resembles the gunman theory of oppression since it assumes that i) women (and men) are interested in maximising their own welfare (in terms of income), and ii) women (and men) have correct beliefs about the oppressive situation.

If Rothstein's model is correct, then (at least part of) the persistent gender inequality in Sweden can be explained by the fact that Swedish women tend to marry older men. The question is, however, whether Rothstein's model can explain everything that we are interested in explaining about gender inequality. Rothstein admits that the model does not show why heterosexual couples tend to display this age difference. He does, however, offer a relatively speculative explanation-sketch of this phenomenon.[1] He assumes that women seek out economically stronger men because they expect to either i) have a weak

[1]Rothstein [2010, p. 15-6]

position on the labour market (due to discrimination), ii) to take the main responsibility of their future children, or iii) both. Since older men tend to be economically stronger, women will seek out older men. Men on the other hand are assumed to either i) have a desire to succeed on the labour market, ii) believe that their partner expects them to be successful, or iii) both. Knowing that their chances to succeed on the labour market goes down if they have to spend time on housework, they will seek out women who they can get to do the housework. Since men believe that younger women fit this description, they will seek out younger women.

According to Rothstein's model, gender inequality is the result of rational and welfare maximising behaviour of individual agents. However, there is nothing in the model that explains why men would have a stronger desire to succeed on the labour market than women in the first place. One possible suggestion is that men are biologically primed to be more competitive than women. Another suggestion is that there exist norms that prescribe how men and women ought to interact with the labour market when they have children. Göran Ahrne and Christine Roman [1997, p. 169] do, for example, argue that the best explanation of gender inequality in Sweden involves norms about motherhood. This is the alternative that we will pursue.

Before moving on, let us say something about norms in general. A norm is usually taken to be a behavioural rule that applies to specific situations.[1] That is, a rule that prescribes certain behaviour in certain situations. A norm of motherhood is a good example of such a rule: if you are a mother you ought to take the main responsibility for your children.

There are many ways a norm can influence a person's behaviour. A mother may for example comply with the norm of motherhood because she does not want to stand out and believes that everyone else follows it. She may also follow it because she believes that transgressions are punished or compliance rewarded. For our purposes, however, we are interested in the mothers who follow it because they have *internalized* it. A person who has internalized a norm will follow it, more or less, unconsciously. The easiest way of explaining what it means to have internalized a norm is probably with the help of an example. Think of a person who has internalized a 'cover your mouth when coughing'-norm. Most of the time when she covers her mouth while coughing, she will not do so because she is afraid of being punished; most of the times she will not even be aware of the norm when she follows it. Rather, she will most often follow it automatically. Since people who have internalized norms do not need any additional motivation, this type of norms seem to be a good candidate for an ideological norm.

[1]For more complete theories of norms, see Allan Gibbard [1990, ch. 4] and Cristina Bicchieri [2006, ch. 1].

This does, of course, not mean that all internalized norms are ideological or oppressive. The 'cover your mouth when coughing'-norm is obviously not ideological. Let us say that a norm is (oppressive and) ideological if it would be the case that if a substantial number of people within an oppressive social order followed the norm, then it would contribute to upholding the oppressive social order. Furthermore, in this section, where we discuss whether norms can explain the persisting gender inequality in Sweden, we are mainly interested in the subset of ideological norms that have been internalized by the members of the oppressed group.

In order to make the concept of internalized norms less mysterious and in order to connect them to the standard Humean model of motivation, let us follow Elinor Ostrom [2005, p. 35] and say that a person who has internalized a norm suffers shame or guilt when she fails to comply. Thus, when a person has internalized a norm she will have her valuation of the associated behaviour changed. For example, if a woman has internalized a norm of motherhood, her attitude towards putting her young children in kindergarten may become unfavourable. Furthermore, a mother who has internalized a norm does not take care of her children because she is aware of the norm requiring her to do so. Rather, under this interpretation, she takes care of her children because not doing so would be associated with guilt and shame.

Ahrne and Roman's suggestion that gender inequality in Sweden is caused by internalized norms of motherhood receives some additional support from the observation that Swedish couples tend to divide household labour relatively equally before they get children. The traditional division of labour, where women take care of the majority of the housework, is introduced after the arrival of the first child. If Swedish men and women have internalized a norm of motherhood, then it could be argued that the first child activates it and causes the parents to assume their traditional gender roles.

Rothstein's [2010, p. 4-5] main objection against explanations in terms of norms rests on methodological grounds. We will postpone the discussion of this objection to chapters 5 and 6. What is at stake here is whether it is possible to explain the persistent gender inequality without norms. We have already established that Rothstein's theory is unable to explain why men have a stronger desire to succeed on the labour market than women. Another problem for the theory can be found in the outcome of more stylised bargaining situations.

Consider for example the experiment described in section 2.5.3. Håkan Holm [2000] carried out an experiment where a group of Swedish undergraduates were paired in twos and asked to play a clash of wills game. The clash of will game shares two features with real bargaining situations.[1] The first is

[1] See, e.g., Schelling [1980, ch. 3].

that the players of a clash of wills game must coordinate with each other or lose altogether. The second is that there exists a conflict of interests since each player prefers the other player to give up a larger share of the limited resources. The payoffs (in SEK) for Holm's game is given by figure 4.7.

		B	
		Hawk	Dove
A	Hawk	$0,0$	$200,100$
	Dove	$100,200$	$0,0$

Figure 4.7: Clash of wills, payoff given in SEK.

In order to avoid making the gender-aspect of the game too salient the experimenters provided the players with information about their co-players' gender with the help of a generic Swedish name. For example, Karl Andersson, if the co-player was male, and Lisa Svenson, if the co-player was female. The players were then asked to split 300 SEK between themselves and their co-player. They were told that if the players agreed on the split both would receive a payoff, otherwise both would go away empty-handed. For example, if player A chose to give 200 SEK to A and 100 SEK to B, and B chose the same distribution, then A would receive 200 SEK and B 100 SEK. If B, on the other hand, chose to give 100 SEK to A and 200 SEK to B, then both A and B would receive 0 SEK each. The players were also asked to play a second game where they had the additional option of splitting the prize equally, i.e., the option of giving 150 SEK to both A and B .

The results of the first game were as follows. When men were paired with men, 55.3% asked for 200 SEK.[1] Likewise when women were paired with women, 55.3% asked for 200 SEK.[2] However, when men were paired with women, 77.1% of the men and 35.7% of the women proposed 200 SEK to the man.[3] In the game when they were allowed to split the money equally, more than 90% of the players chose to do so.

At first sight the results seem to support Ahrne and Roman's [1997] explanation in terms of internalized norms. In fact, the experiment seems to provide evidence for two lexically ordered norms. One norm that regulates the distribution of resources if an equal split is available, and one that applies to situations where an equal split is unavailable and where the opponent is believed to be of a different gender than oneself.

However, as Holm [2000, p. 304] points out, information about gender allows the players to coordinate on the Pareto-optimal outcome. After all,

[1] 26 out of 48.
[2] 21 out of 38.
[3] 37 out of 48 men, and 10 out of 28 women asked for 200 SEK.

same-sex pairs successfully coordinate their behaviour in about a half of the games, whereas different-sex pairs succeed more often. It could, therefore, be argued that female players do not play dove because they believe that they ought to do so, rather they play dove because they expect men who play against women to play hawk. On this reading, gender works as a focal point. Although the existence of the focal point can be explained in terms of internalized norms, it is also possible to explain it in terms of past play and rational expectations. Assume that a female player has experienced a large number of clash of wills games in the past. Furthermore, assume that in the majority of the games against male players, the male player played hawk. If she lacks any other kind of information, it seems rational for her to expect her male opponent to play hawk in this game and thus for her to play dove.

The result of Holm's experiment seems to lend equal support to both the rational action and the ideological norm hypotheses. After all, if a female player is interested in maximising her interest-satisfaction (and if doing so implies maximising monetary payoff), and has correct beliefs about her male opponents, then she will propose 200 SEK to her male co-players. Similarly, if a female player has internalized the norm that men ought to have more when limited resources are divided, then she will propose 200 SEK to her male co-players. Thus, we are unable to draw any definite conclusions about whether the outcome was caused by rational interest-maximising behaviour or ideological norms.

Although Holm's experiment does not allow us to draw any conclusions about the necessity of ideological norms to explain gender inequality, it points us in the direction of how to construct a method for doing so. Cristina Bicchieri [2008] uses a more sophisticated version of the clash of wills game, a so-called ultimatum game, to investigate the existence of fair division norms. Although Bicchieri's social norm framework is much richer than the one we employ, we can still use her basic method to determine whether we have reason to believe in the existence of gender-based division norms.

An ultimatum game is played by two players: a proposer and a respondent. The proposer is given Y resources and asked to offer a share, $0 \leq x \leq 1$, to the respondent and keep the rest. After she has heard the proposer's offer, the respondent gets the choice of accepting or refusing the offer. If accepted the respondent receives xY resources, and the proposer receives $(1-x)Y$ resources.

A rational respondent who is solely concerned with maximising expected monetary value will accept any offer as long as it is positive.[1] A rational proposer who believes that the respondent is solely concerned with maximising expected monetary value, will therefore offer the smallest possible positive

[1] If $x = 0$ then this type of respondent will be indifferent between accepting and refusing.

share ($x > 0$). In experimental situations, however, players do not behave like this, rather the proposers tend to make offers $0.4 \leq x \leq 0.5$ and the respondent tend to refuse offers where $x \leq 0.2$.[1] Ernst Fehr and Klaus Schmidt [1999] have attributed this to a preference for fair allocations. We could say that the result indicates that people tend to have an internalized norm of fairness.

One of the important differences between Holm's clash of wills game and the ultimatum game is that the latter allows for a more diverse bundle of offers. This allows us to both identify and measure the strength of a potential gender-based division norm. If there exists a gender-based division norm that applies to the division of resources in the ultimatum game, then we would expect a woman in a mixed-sex pair to both offer greater xs and accept lower xs, than a woman in a same-sex pair. Conversely, a male player in a mixed-sex pair would also be expected to offer smaller xs and refuse greater xs, than a man in a same-sex pair. The experiment also allows us to measure the strength of the gender-based division norm by comparing the results with the results of a game where information about gender is not provided. The same measure would allow us to compare the strength of gender-based division norms across cultures and societies. Furthermore, if a gender-based division norm regulates the behaviour of players in the ultimatum game, then it is difficult to argue that this norm is beneficial for women as well as for men. After all, knowledge of gender will not increase the expected payoff for women, as it did in Holm's clash of wills game, rather it will, if a norm exists, ensure that women receive less both as proposers and respondents.

Sara Solnick [2001] has used a version of the ultimatum game to investigate whether knowledge about gender affects the outcome. Undergraduates were divided into two groups and assigned roles of proposers and respondents. A subset of the pairs received information about their co-players first name, whereas the others did not. They were then asked to play the ultimatum game with \$10.[2]

On average both male and female proposers made less generous offers to female respondents than to male respondents when gender was known. Female respondents were on average offered \$4.37 whereas male respondents were on average offered \$4.89. Of all the offers made to male respondents, 82% were

[1] Colin Camerer [2003, p. 49] has collected the results from 15 studies of ultimatum games in experimental settings. He reports that in these studies, the median and modal ultimatum offers were usually 40%-50%, and the mean offers were between 30% and 40%. Furthermore, he reports that offers between 40% and 50% are rarely rejected, whereas offers below 20% are rejected about half the time.

[2] One difference between the ultimatum game used by Solnick and the standard ultimatum game was that in her experiment the respondents were asked to write down their minimum acceptable offer before they had received the proposers offer.

an even split or greater (the proposer offers $x \geq 0.5$), as compared to 59% of the offers to female respondents. Furthermore, respondents (of both sexes) tend to demand a higher minimum acceptance offer from females than from males: \$2.81 from males and \$3.42 from females. These results lend some support to the claim that there exists a gender-based division norm stating that females ought to give up more in bargaining situations. Furthermore, if the norms explain the outcome of the stylised bargaining game, then it could be argued that they can also be used to explain the existence of gender inequality in general.

It should, however, be mentioned that, in order to make sense of all of Solnick's results, a richer theory of norms would be needed.[1] We would, for example, have to distinguish between players who follow norms because they are internalized, players who follow norms because others follow them, and players who follow norms because they fear punishment. We would also need a theory of how different norms become activated in different situations,[2] in order to investigate whether potential chivalry norms (giving more to women because they are women) interfere with the results.[3]

For our purposes, however, it is enough to get some results indicating that we need internalized norms to offer a satisfactory explanation. It is, however, also interesting to note that there exists a method that allows us to both investigate the existence of ideological norms in a society, and to compare different societies with respect to the strength of these norms.

Finally, let us point out that it is possible for ideology to consist of either false beliefs or internalized norms, or of both. In the case of the persistent economic inequality in the US, there is no need to attribute ideological norms to America's poor. In order to explain the persistence of the inequality, it is enough to point to their false beliefs that their society has a high degree of social mobility. Similarly, an explanation that involves a gender-based division norm does not need to involve false beliefs. It is, for example, very likely that Swedish mothers are aware that they take the main responsibility for the children and earn less than their male counterparts. The reason that gender inequality persists, according to the ideology theory of oppression, is because Swedes have internalized a norm of motherhood. Finally, explanations of oppression in terms of religion often refer to both false beliefs and internalized norms. Religious people are usually attributed a (false) belief about the existence of an after-life. This belief may cause them to devote themselves to their religion in the hope of a better reward than any post-revolutionary social order

[1] For example, Bicchieri [2006, 2008].

[2] See, e.g., Tversky and Kahneman [1981] and Camerer and Thaler [1995].

[3] See, e.g., Eckel and Grossman [2001].

could give them.[1] In addition to being victims of false beliefs, they are usually thought of as having internalized norms that can prevent them from engaging in revolutionary activity. The most famous example is probably Christianity's appeal to turn the other cheek.

4.3 Explanatory importance of ideology

So far we have shown that the gunman theory of oppression has problems explaining some of the interesting aspects of the persisting economic inequality in the USA and gender inequality in Sweden. In the former case, we saw that although it was possible to construct a model that showed how inequality persists in a democracy, we were not provided any reasons to believe that the situation in the USA is accurately described by the model's assumptions. Furthermore, the model failed to account for the Americans' false belief that there is a high degree of social mobility in the USA. In the latter case we examined a model that attempted to explain gender inequality in Sweden. This explanation, however, failed to explain why men have a stronger desire than women to succeed on the labour market. Moreover, it could not provide a straightforward explanation of gender inequality in the more stylised bargaining situations provided by clash of wills and ultimatum games. We also showed that the ideology theory of oppression can provide relatively straightforward explanations of these phenomena with the help of ideological beliefs and norms.

Now that we have some reasons to believe that ideology is needed to explain some of the persistent oppressive social orders, let us formulate the ideology theory of oppression. Recall that we formulated the mystery of oppression as follows:

> THE MYSTERY OF OPPRESSION: There exists a persistent social order, O_1, a possible persistent social order, $O_2 \neq O_1$, and a group X, such that
>
> 1. X is more oppressed in O_1 than in O_2,
> 2. O_2 is better than O_1 (in terms of welfare) for the members of X, and
> 3. X has the ability to exchange O_1 for O_2.

Also recall that the gunman theory of oppression explained the mystery by denying condition 3. It showed that since the members of X were subject to

[1] In all fairness, this belief may also cause them to engage in revolutionary activity since it allows them to perceive the cost of detonating bombs strapped to their chests to be low relative to the reward.

a free-rider problem X did not have the ability to exchange O_1 for O_2 in any relevant sense.

The ideology theory of oppression explains the mystery by pointing out that even if conditions 1-3 are fulfilled, the members of X will not revolt unless they realise that the conditions are fulfilled. Holding false beliefs can prevent them from realising that 1 or 3 are fulfilled. For example, the false belief that American society has more social mobility than it actually has, is a belief that masks the truth of condition 1. To see this, remember that we, in section 2.4, argued that unequal social orders with a low degree of social mobility are oppressive. Moreover, we argued that under normal circumstances people tend to dislike social orders that are oppressive in this sense. So if a poor American is convinced that the American dream is alive and well in the US, then she will not perceive American society as oppressive, and consequently she will not be motivated to exchange it for another order.

Furthermore, although O_2 is better than O_1 for the members of X this does not mean that the members of X will prefer O_2 to O_1. Having internalized a norm that social order O_1 ought to be the case is an effective way of causing the members of X to prefer O_1 to O_2. For example, the gender-based division norm causes women to prefer distributions of resources that give them less. Although women may realise that they are oppressed and that they have the ability to exchange the present social order for a less oppressive order, they may not do so because the internalization of a gender-based division norm causes them to prefer an unequal distribution of resources to a more equal distribution.

Let us collect these types of beliefs and norms under the heading of *ideology*. We can now say that, according to the ideology theory of oppression, oppressive social orders persist because their members are subject to ideology.

It is perhaps possible for a proponent of the ideology theory of oppression to stop here and be satisfied with this explanation of the mystery of oppression. That ideology is responsible for the lack of civil unrest is an interesting find in itself. It tells us, among other things, that if we want to mobilise an oppressed group against an oppressive social order, then we should start by getting rid of their ideological beliefs and norms. After all, before the pacifying ideology is gone, the oppressed will not even realise that there is a problem that needs to be solved. In a sense, finding out that the reason that people do not engage in revolutionary activity is that they are mystified is analogous to finding out that the reason people are dying is that the water supply is poisoned. We do not need to know why there is poison in the water in order to issue a warning. Similarly, we do not need to know the why and how of their mystification in order to know that we need to do something about it. However, just as it would be good to know how the poison got into the water supply in order to prevent it

from happening again, it would be good to know why the citizens are mystified in order to do find out the best means of getting rid of their ideology.

A more serious objection to a version of the ideology theory of oppression that remains silent about why people become victims of ideology, is that it leaves something important out of the picture. Once we have been told that Americans refrain from revolutionary activities because they falsely believe that their social order has a higher social mobility than it actually has, a new pressing question arises. Why do Americans hold a belief that goes against their interests? Although we could just assume that people have false beliefs this would hardly be satisfactory. For the same reasons that we struggled to show that selective incentives (in section 3.4.1) could arise within the framework of the gunman theory of oppression, we should try to expand the ideology theory of oppression so that it can provide an explanation of why people suffer from ideology.

According to the most common version of the ideology theory of oppression the existence of ideology is explained in terms of the consequences or function of the ideology. This is the version that, e.g., Michael Rosen [1996] ascribes to Marx and Cohen, and it is the version that will be defended in the following chapters. We will postpone the discussion about the problem of functional explanation in general and potential underlying mechanisms to chapters 5 and 6 respectively. The remainder of this chapter will be dedicated to formulating the functional component of the ideology theory of oppression and taking care of some preliminary objections.

Let us begin with Rosen's formulation of the ideology theory of oppression:

> illegitimate societies [...] maintain themselves without depending solely on coercion [...]
>
> 1. in virtue of false consciousness in the part of the citizens, [and]
> 2. this false consciousness occurs in response to the needs of society. [Rosen, 1996, p. 260]

Although this formulation captures the gist of the ideology theory of oppression, it is flawed for at least two reasons. First, as claim 1 is stated it seems as the ideology theory requires that *all* citizens are victims of ideology (or false consciousness as he calls it). As we have seen, however, it is enough that some of the citizens are victims of ideology in order for the oppressive status quo to be upheld without depending solely on coercion.

The second problem with the formulation is that claim 2 suggests that the ideology theory of oppression is committed to explaining not only why the

members of the oppressed group, X, are victims of the ideology I, but also how I came about in the first place. If this is correct, then the ideology theory of oppression has to show how the needs of society produced the first ideology-token. That is, it would have to provide the causal connection between the needs of society and the thoughts of the individual who first came to hold the false belief.

However, a more plausible version of the ideology theory of oppression (e.g., Cohen [2000]) does not make any claims about how the first token of an idea came about. It is content with showing that if an idea serves the function of upholding an oppressive status quo is present in a population, then this idea will remain (and spread) in the population for as long as it serves this function. This can be compared to functional explanations in evolutionary biology where it is enough to know that a trait that maximises fitness is present in a population in order to explain that it will remain or spread because it is fitness maximising.[1]

In order to avoid these implication let us reformulate the ideology theory of oppression as follows:

> IDEOLOGY THEORY OF OPPRESSION:
> Oppressive societies maintain themselves without depending solely on coercion
>
> 1. in virtue of ideological beliefs and norms among a substantial part of the citizenry, and
> 2. these citizens are subject to ideology because this serves the function of upholding the oppressive status quo.

Now that we have formulated the ideology theory of oppression we can turn to some preliminary objections. Both Elster [1983b] and Tännsjö [2006] admit that members of oppressed majorities are victims of ideological beliefs in the sense that they, e.g., falsely believe that social mobility is higher than it actually is. They do, however, argue that both the ideological beliefs and the oppressive situation is caused by a third factor, namely coordination problems of the type described in chapter 3. Tännsjö [2006, p. 429] argues that

> even if revolution is in principle possible and collectively desirable, each individual is not willing to contribute his or her share to the common endeavour. When this is the case, then it is only to be expected that all sorts of ideological excuses should pop up. But it is not ideology as such which explains the stability of the

[1] We will return to this in chapter 5.

> oppressive system. It is rather the coordination problem that explains both the stability of the system and the emergence of the ideology. This is what has gone unnoticed to the ideology theorist.

There are two claims here that need to be addressed. The first, directed at condition 2, is that ideological beliefs do not arise in response to the needs of society, but rather as a means of minimising cognitive dissonance. The second, directed at 1, is that since both ideology and the stability of an oppressive social order is caused by a third factor (a coordination problem), ideology can play no part in explaining the persistence of the social order. If Tännsjö is right, then ideology will become a mere byproduct of oppression.

Concerning the first claim, if the ideology theory had claimed that ideological beliefs occur in response to the needs of society, as Rosen's formulation implies, then Tännsjö's claim would be a refutation of the ideology theory of oppression. As we mentioned above, however, the ideology theory of oppression does not say anything about how the first token of an ideological belief came about. The claim that the first instance of the American dream came about in order to minimise the cognitive dissonance of some of America's poorer citizens is perfectly compatible with the ideology theory of oppression' claim that the American public will continue to believe in the American dream for as long as it serves the function of preserving the oppressive social order.

Furthermore, the ideology theory of oppression is silent about the mechanism that preserves and spreads ideology. All it says is that the oppressed believe in the American dream because this belief serves the function of preserving the oppressive status quo. This can be compared to the claim that car factories are large because it serves the function of maximising profits. This functional claim is compatible with both, for example, i) a mechanism of purposive action where the managers expand the factories in order to maximize profits, and ii) a Darwinian mechanism where car factories that fail to maximize profits go extinct. It is not entirely clear that a discovery of the underlying mechanism will eliminate the explanatory power of the functional explanation of the car factory size. Similarly, even if it were shown that ideology spreads through a mechanism of minimising cognitive dissonance it is unclear whether this would show that the ideology theory of oppression was false.[1]

Finally, that people tend to minimise cognitive dissonance does not necessarily give us reason to believe that ideology is produced by a coordination problem. There are, after all, two ways to deal with cognitive dissonance. The first is to have one's beliefs or preferences altered so that they suit reality, the second is to try to change the world so that it suits one's beliefs and prefer-

[1] More about this in chapter 5.

ences. In violently repressive societies, the first might be the natural way to deal with cognitive dissonance; in Western democracies, on the other hand, it seems easier to join a political movement and try to change the world. Thus, knowing that people tend to minimise cognitive dissonance might add to our surprise when we realise that people do not engage in revolutionary action. It could therefore be argued that evidence for a psychological mechanism that minimises cognitive dissonance gives us all the more reason to suspect the existence of ideological beliefs and norms.

Let us turn to the second claim that stated that since both ideology and oppression are caused by a third factor, ideology cannot help explain the persistence of oppression. Let us first note that there seems to be no reason to believe that if A has caused both B and C, then B cannot explain C. At least not if the events are spread out in time. Assume that Fidel lifts a log and then positions a rock under it so that the log remains in a lifted position. Although initially Fidel was responsible for both the log's and stone's position, at a later time it is the stone that keeps the log in place. Or in our case, the fact that a coordination problem, at time t_1, causes both the oppressive social order and ideological beliefs, does not mean that the ideological beliefs are not responsible for preserving the oppressive social order at a later time t_2.

Furthermore, if we abandon the idea that an explanation of oppression must be an all or nothing affair, then there are reason to believe that both the difficulty to solve coordination problems and ideological beliefs can play a part in explaining the persistence of oppressive social orders. Even if a coordination problem was enough to preserve an oppressive social order, S, by itself, this would not imply that ideological beliefs cannot play a role in the explanation as well. In terms of our previously described game, let us assume that the n-player assurance game shown in figure 4.6 correctly describes society S. The failure to solve the coordination problem provides one explanation of the lack of revolutionary action in this society. However, assume that $n-k+1$ people are deluded by ideological beliefs. This means that ideology provides a second explanation. If this is the case, then the lack of civil unrest is overdetermined by the coordination problem and the ideological beliefs.

Next, assume that some but less than $n-k+1$ people in S are deluded. In this case, ideology is not by itself enough to explain why the social order persists. After all, the revolution will succeed if everyone who is not deluded engages in revolutionary action. However, this does not mean that ideology does not help explain the lack of a civil unrest. This situation can be compared to Che and Fidel's attempt to push a car up a hill. Fidel is much stronger than Che, and can push the car all by himself, whereas Che cannot do so. Che does, however, join in and help. Just as leaving out Che's effort would be wrong when explaining that the car was pushed over the hill, so would it be wrong to

leave out the ideological beliefs when explaining the lack of revolution. After all, the more oppressed that are convinced of, e.g., the American dream, the smaller the number of potential revolutionaries become, and consequently the harder it gets to organise a successful revolution.

4.4 Summary

In this chapter, we have provided an argument for the explanatory necessity of ideology along the lines of Durkheim's famous argument for the existence of social entities. However, instead of conducting a thorough empirical investigation of European suicides, we remained in our armchair and investigated whether some versions of the gunman theory of oppression could explain the persistent inequalities in the USA and Sweden. We concluded that since this could not be done, and since the ideology theory of oppression could provide a relatively straightforward explanation of these cases, there are some reasons to believe that ideological beliefs and norms are responsible for the persistence of oppressive social orders.

Once we had offered this argument, we moved on to formulate the version of the ideology theory of oppression that will be defended in this book. In addition to the claim that oppressive social order persist because their members suffer from ideological beliefs and norms, it was claimed that they suffer from ideological beliefs and norms because this serves the function of upholding the oppressive social order. Although we will dedicate the next two chapters to a defence of the functional claim we took care of some preliminary objections. We did for example point out that the objection that ideological beliefs are best explained with the help of psychological mechanisms does not (by itself) refute the ideology theory of oppression. We also showed that it is a mistake to conclude that ideology cannot play a role in explaining continued oppression if it is shown that a coordination problem is initially responsible for both the oppressive social order and the ideological beliefs.

In the following two chapters we will try to make sense of the functional component in the ideology of oppression. The next chapter will investigate whether functional explanations can ever be acceptable in the social sciences. In chapter 6 we will investigate whether there are any micro-mechanisms describing belief-formation compatible with the ideology theory of oppression that do not presuppose ontologically queer entities.

5. On rain dances, mechanisms, and functional explanations

5.1 Introduction

In this chapter, we will focus on whether functional explanations can ever be appropriate in the social sciences. In our investigation we will use the explanation of the Hopi's rain dance as a test case. The discussion of whether the functional claim of the ideology theory of oppression is valid will be postponed to chapter 6.

So, why did the Hopi perform their rain dance? If asked they would probably have answered that they performed it in order to produce rain. Although this might have been what the Hopi believed, some have argued that the real reason they danced was because it increased social cohesion in Hopi society.[1] The latter is an example of a functional explanation as it seeks to explain the rain dance in terms of its function to increase social cohesion.

Many have been suspicious of functional explanations. It is sometimes claimed that they postulate effects that precede their causes. For example, the performance of the rain dance is explained by the increase of social cohesion that follows the performance of the rain dance. It is concluded that since causes must precede their effects, functional explanations are seriously flawed.

It has also been objected that functional explanations do not comply with the standard deductive-nomological (DN) model of explanation. It has, for example, been claimed that it is impossible to deduce the performance of a rain dance from an explanans containing propositions about the function of the rain dance and initial conditions of Hopi society.[2]

Finally, it has been argued that functional explanations in the social sciences lack access to the type of mechanism that vindicates the use of functional explanations in biology. For example, when evolutionary biologists explain the giraffe's long neck by citing its function to allow the giraffe to reach leaves high up in the tree, they implicitly refer to the underlying mechanisms

[1] See, for example, Robert Merton [1968].

[2] We will return to this in more detail when we look at Hempel's argument against functional explanations below.

provided by genetics and natural selection. Social scientists, on the other hand, do not seem to have a similarly well-confirmed mechanism that connects the function of the rain dance with its performance. Since social scientists cannot provide this type of mechanism, it seems as they cannot justify their use of functional explanations.

The last objection is usually associated with methodological individualism. Jon Elster [1980] has, for example, argued that Gerald Cohen's [1978] appeal to functional explanations in his defence of Marx's historical materialism failed because he did not provide an underlying mechanism. Elster's argument was dubbed the *missing mechanism objection*.[1] The objection lead to an extensive discussion of whether mechanisms were necessary for functional explanations.[2] After the mid-90s the interest in the appropriateness of functional explanations in the social sciences slowly faded. However, as it did so, the interest in mechanisms increased.[3] In other words, since Elster's original formulation of the missing mechanism objection a lot of research has been dedicated to the study of mechanisms.

The main purpose of this chapter is to defend the use of functional explanations in the social sciences against the missing mechanism objection. We will look at two different versions of the objection that both rest on the claim that functional explanation in the social sciences are unacceptable if an underlying mechanism is lacking. The first is expressed in the form of a dilemma: if a mechanism is not provided, then functional explanations commit us to the existence of ontologically extravagant social forces; if, on the other hand, a mechanism is provided then the functional explanation will become redundant. This will be the topic of section 5.4. In section 5.5, we will look at the second argument that rests on epistemic considerations and claims that we are never justified in accepting a functional explanation unless an underlying mechanism has been provided.

However, before we study the missing mechanism objections we will, in section 5.2, spell out the functional explanation of the Hopi rain dance in detail. This will be done with the help of Cohen's [1978] account of functional explanations in terms of consequence laws. Once we have done so, we will be able to take care of two preliminary objections: that functional explanations imply that the effect precedes the cause, and that functional explanation do not comply with the DN-model of explanation. We will, in section 5.3, say some-

[1] See, e.g., Philip Pettit [1996].

[2] See, e.g., the exchange between Elster, Cohen, and others in *Theory and Society* 1982.

[3] In the philosophy of science in general see, e.g., Bunge [1997], Machamer et al. [2000], Glennan [2002], and in the social sciences in particular Hedström and Swedberg [1996] as well as Hedström and Ylikoski [2010].

thing more general about what mechanisms are and how they relate to Cohen's account of functional explanations.

It will be argued that recent developments in the social mechanism literature show that it is relatively straightforward to reformulate Cohen's account of functional explanations in terms of a (macro-)mechanism. This indicates that the problem with functional explanations cannot be that they fail to provide mechanisms. Rather, it seems as the objection should be reformulated so that it states that functional explanations in the social sciences are inappropriate because they fail to provide the *right* type of mechanism. We could call this the *wrong-kind-of-mechanism objection*. When scrutinised the proponents of the objection seem to think that the problem with functional explanations is that it has not provided a micro-mechanism in terms of individuals and their properties. Once this has become clear, the question of whether functional explanations are acceptable will depend on whether explanations in terms of social facts are acceptable.

5.2 Functional explanations with consequence laws

In his defence of Marx's historical materialism, Cohen [1978] had to make sense of functional explanations. At the time, the most common objection against functional explanations was that they do not comply with the DN-model of scientific explanation. According to the DN-model an explanation is divided into an explanandum and an explanans.[1] The explanandum consists of a proposition that describes the phenomenon that is to be explained. For example, the mercury in the barometer rises. Whereas the explanans consists of the set of propositions that are supposed to explain the explanandum. The explanans can be further divided into two subsets: a set of initial conditions, and a set of lawlike connections. For example, the initial condition that air pressure increases and a lawlike statement that connects changes in air pressure with the behaviour of the barometer.

According to Hempel, all propositions in the explanans have to be true in order for it to explain the explanandum. This means that both the initial conditions and the lawlike statements must be true. That the lawlike statement is true is usually taken to imply that it constitutes a natural law, or that it is derivable from some set of natural laws and initial conditions. For example, although the generalisation that "whenever air pressure increases, the mercury in the barometer rises" does not instantiate a natural law, it is (probably) derivable from some set of natural laws and initial conditions.

It might seem too much to demand that the explanans be true. After all,

[1] See, e.g., Hempel [1965, p. 247].

if science is fallible, how can we ever know when we have hit upon a natural law? Would it not suffice for the explanans to be highly confirmed by all available and relevant evidence in order for the explanans to be able to explain the explanandum? Hempel considers this alternative, but dismisses it since it would put us in an awkward position with respect to earlier stages of science. If, at an earlier stage, a phenomenon was claimed to be explained by an explanans that was well confirmed by the evidence available at the time, then we would have to say that it was actually explained even if recent findings show that the explanans was false. Since it sounds strange to say that an explanation was correct at one time but not at another, Hempel concludes that we should demand that the explanans is true.

However, recall that we in section 3.5 argued that models with false assumptions could be explanatory. Within the framework of Hempel's DN-model, we would have to say that the false assumptions of the gunman theory of oppression describe a possible model world. Since the assumptions are true in the model world, we are able to deduce the explanandum from a true explanans in the model world. The explanatory power of the model with respect to phenomena in the actual world, depends on how credible the model world is. That is, the closer the model world is to the actual world, the higher will its explanatory power be with respect to the actual world.[1]

What is mainly at stake for us, however, is not whether the explanans of a functional explanation is true, but rather whether we can have equally good *reasons to believe* that a functional and, e.g., a rational choice explanation in the social sciences are true. That is, if our functional explanations make false assumptions, then this will not be a problem unless it will render the model world (much) less credible than the model worlds of rational choice explanations.

Before we turn to functional explanations, let us return to the DN-model and note that if the propositions in explanans are true, then the explanans is said to explain an explanandum if the explanandum can be logically deduced from the propositions in the explanans. Consider, for example, the following explanation of the behaviour of a barometer:

I. Whenever air pressure increases, the mercury in the barometer rises.

II. Air pressure increases.

III. Therefore, the mercury in the barometer rises.

Premise I describes the lawlike statement that connects air pressure to the behaviour of the barometer, and premise II describes the empirical observation

[1] We will return to this below.

that air pressures increases. This allows us to deduce and, if the explanans is true, explain the rise of the mercury in the barometer.

With respect to functional explanations, Hempel [1965, p. 310] argued that they fail to comply with the formal requirements of the DN-model of scientific explanation. Along the lines of Hempel's discussion, the functional explanation of the Hopi rain dance should be spelled out as follows:

1. Hopi society functions adequately,

2. Hopi society functions adequately only if it has a high degree of social cohesion, and

3. if rain dances are regularly performed in Hopi society, then (as an effect) it will have a high degree of social cohesion.

4. Therefore, rain dances are regularly performed in Hopi society.

As the argument stands, it is invalid. To be more precise it is invalid since it commits the fallacy of affirming the consequent with respect to premise 3.[1] We cannot infer that the Hopi perform rain dances from the claim that rain dances lead to high social cohesion and that Hopi society has high social cohesion. Doing so is equivalent to inferring that all cats are black from the following two premises: 1) if all cats are black, then Findus is black, and 2) Findus is black. Since the explanandum cannot be deduced from the premises, Hempel concluded that functional explanations are seriously flawed. Consequently, we do not have to investigate whether the propositions in the explanans are true in order to dismiss the functional explanation of the Hopi's rain dance.

However, consider the following change of premise 3:

3. (b) Only if rain dances are regularly performed in Hopi society, will Hopi society (as an effect) have a high degree of social cohesion.

If we replace premise 3 with 3b, it will become possible to deduce the explanandum from the new explanans.[2] From premises 1 and 2 we can deduce that Hopi society has a high degree of social cohesion. According to premise

[1] This can be easily be seen in the following formalisation of the argument:

1. A
2. $A \rightarrow B$
3. $C \rightarrow B$
4. C.

[2] Formally, this is equivalent to reversing the arrow in the third premise:

3b the only way of achieving a high degree of social cohesion in Hopi society is by regularly performing rain dances. Therefore, if Hopi society has a high degree of social cohesion, then rain dances must be regularly performed in Hopi society.

The problem, however, according to Hempel [1965, p. 311], is that premise 3b is empirically questionable. After all, there seems to be many other activities, that do not involve rain dances, that can increase social cohesion. For example, organised sports, drama, or other types of ritualistic dances.

At most, Hempel argues, it is possible to explain that the rain dance *or* some functional equivalent ritual exists. For example, let I be the set of functional equivalent rituals any of which can ensure that Hopi society has high social cohesion. It is then possible to use this definition and premise 1 and 2 to make the following deduction of a new explanandum:

1. Hopi society functions adequately,

2. Hopi society functions adequately only if it has a high degree of social cohesion, and

3. (c) only if one of the rituals in I is regularly performed in Hopi society, will Hopi society (as an effect) have a high degree of social cohesion.

4. (c) Therefore, one of the rituals in I is regularly performed in Hopi society.

Although the inference of 4c is valid, and premise 3c is not empirically questionable, Hempel [1965, p. 314] believes that inferences of this type are rather trivial. After all, instead of explaining why a rain dance is performed we now have an explanation of why some activity that increases social cohesion is performed. We will return to the functional equivalence objection below.

As the functional explanation (1-4) of the Hopi's rain dance stands, Hempel is right that it fails to deduce the explanandum. One way of saving the functional explanation would be to deny that the DN-model is the best model of scientific explanation, and then show that according to the correct model functional explanations are not formally flawed. Although it is easy to point to problems in the DN-model, it is not as easy to find an alternative that has the

1. A
2. $A \rightarrow B$
3. $C \leftarrow B$
4. C.

same immediate intuitive appeal. It would, therefore, be desirable if we could show that, contrary to Hempel, functional explanations do not violate the requirements of the DN-model. Cohen provides us with an alternative way of explicating functional explanations within the DN-model's framework.

Cohen begins by observing that in some contexts, statements of the type "f preceded e" are considered to be explanatory because they are supported by the generalisation that whenever F occurs, E occurs. He then asks whether it is possible to find a similar generalisation to justify function-statements of the type "(token) e occurred because of its propensity to cause (events of type) F."[1]

He suggests that what makes functional statements explanatory is that they rest on a so-called *consequence law*. A consequence law connects the disposition of a system to produce a (usually beneficial) effect whenever a certain type of event occurs, with later occurrences of events of this type. For example, in the case of the Hopi's rain dance, the law would connect the disposition of Hopi society to increase social cohesion (the beneficial effect) whenever a rain dance is performed, with the performance of rain dances at later times. According to Cohen, the explanatory power of the statement "e occurred because of its propensity to cause F" is justified by the generalisation that whenever E would cause F, E occurs.

The general form of a consequence law is the following:

1. IF it is the case that *if* an event of type E were to occur at t_1, *then* it would bring about an event of type F at t_2, THEN an event of type E occurs at t_3. [Cohen, 1978, p. 260]

Although the nested conditionals lend the consequence law an awkward look it becomes easier to digest if we think of the minor conditional (the conditional starting with an italicised and small '*if*') as describing a disposition of a system. This gives the consequence law the same form as other lawlike statements. For example, the lawlike statement that connects increased air pressure, A, to the rise of the mercury in the barometer, B, can be written 'if A then B.' If we use D as an abbreviation for the disposition, then the consequence law can be written 'IF D THEN E.' For example, if Hopi society has a disposition to increase social cohesion when a rain dance is performed, then rain dances will be performed.

The disposition expressed by the minor conditional should be interpreted as a statement about something, e.g., a society, an institution, an organism, etc. To say that a society has a disposition of this type is to say that were an event

[1]Let us use small letter to signify event-tokens and capital letters to signify event-types, so that e is an event-token of the event-type E.

of type E to occur in this society, then, as an effect, an event of type F would follow. For example, to say that Hopi society has a disposition to increase social cohesion when a rain dance is performed, is to say that if the Hopi were to perform a rain dance, then social cohesion would increase.

Note that the focus on dispositions alleviates the worry that functional explanations postulate that the effects precede their causes. After all, according to this interpretation, the cause of a rain dance is not the increased social cohesion that follows its performance, but rather *the disposition* of Hopi society to increase social cohesion whenever a rain dance is performed.

Furthermore, it is worth pointing out that although the consequence law is stated as deterministically, it can also be interpreted in probabilistic terms. That is, that a society has a disposition to produce an event of type F when E occurs, *increases the probability* that E will occur at a later time. As [Cohen, 2000, p. 262] points out, this does not set consequence laws apart from normal causal laws. For example, in order to explain that a person who has just had four cups of coffee has trouble sleeping, we could refer to the lawlike generalisation that "whenever someone drinks four cups of coffee she is sleepless." Although the statement is false in the sense that there are exceptions to it, we would accept it as part of an explanation since it is true that most people become sleepless when they drink four cups of coffee. For the sake of simplicity, however, we will talk of laws as being deterministic.

How does Hempel's objections fare against Cohen's explication of functional explanations. In order to see this, let us first formulate the functional explanation of the Hopi's rain dance as follows:

L IF it is the case that *if* a rain dance were to be performed, *then* social cohesion would increase, THEN a rain dance is performed.

C In hopi society, *if* a rain dance were to be performed, *then* social cohesion would increase.

E Thus, a rain dance is performed in Hopi society.

Concerning Hempel's formal objection that we cannot deduce the explanandum from the explanans, it should be clear that Cohen's use of consequence laws allows us to deduce **E** from **L** and **C**. The objection that there are functional equivalents, however, cannot be as easily avoided. After all, the existence of functional equivalents do not disappear just because we reformulate the functional explanation.

In order to investigate the problem of functional equivalents, let us assume that we have discovered other activities that increase social cohesion among other groups. We might, for example, discover that among the Sioux, a ghost dance increases social cohesion, or that among Europeans, organised sport and

religion serves the same function. However, since the consequence law, **L**, does not say anything about a rain dance being the only means of increasing social cohesion, the discovery of functional equivalents does not contradict **L**. What the discovery might do, is to count as evidence against this consequence law. After all, the discovery of functional equivalents gives rise to a new question: why should the rain dance's potential suffice for its actualisation when other ceremonies with similar potential are not actualised? [Cohen, 1978, p. 274]

Cohen provides three lines of defence against this objection. First, although some alternative ritual increases social cohesion in some distant society, this does not mean that this ritual would increase social cohesion among the Hopi. There may, for example, be external factors that make the rain dance better suited for the Hopi and the ghost dance better suited for the Sioux. For example, since the subsistence of Hopi society is dependent on agriculture for its survival it seems as a rain dance lies closer at hand than, e.g., a ghost dance. Among the Sioux, on the other hand, who were mainly hunters and therefore not as dependent on rain as the Hopi, a ghost dance might have been easier to motivate.

Second, even if it was shown that no external factors made the performance of a rain dance better suited than the ghost dance, there could still be good reasons to believe that a rain dance (and not a ghost dance) would be performed. After all, just because another ritual has the same potential to increase social cohesion in Hopi society, it seems unlikely that it will be performed if it is not a part of the Hopi's *traditional repertoire*. That is, if the Hopi are not accustomed to the ghost dance, then it might be as difficult for them to become excited about a ghost dance as it is for Europeans to become excited about American football. The reason why the rain dance is part of the Hopi's cultural repertoire may be a historical fluke. The Hopi could as well have performed a ghost dance, but, for some (or no) reason they started to dance a rain dance instead and have been doing so ever since.

Cohen [2000, p. 275] argues, that the discovery of functional equivalent of this type will not give us a reason to completely reject an explanation of **E** with the help of **L** and **C**. Instead, it gives us a reason to modify the lawlike statement and add a premise stating that the rain dance is part of the traditional repertoire:

L' IF it is the case that *if* a rain dance were to be performed, *then* social cohesion would rise shortly thereafter, *and* rain dancing belongs to the traditional repertoire, THEN a rain dance is performed.

C *If* a rain dance were to be performed, *then* social cohesion would rise shortly thereafter.

C' Rain dancing belongs to the traditional repertoire.

E Thus, a rain dance is performed.

Third, even if there were other dances in the traditional repertoire, it is still possible to relax the explanandum in the way that Hempel suggested. Instead of explaining the performance of a rain dance, we would have to be satisfied with explaining the performance of some ritual. Whether such results are as trivial as Hempel argues, is a matter of what we are interested in. If we are interested in why a rain dance rather than ghost dance is performed, then this would indeed be a rather trivial answer. However, if we are interested in why the Hopi perform a rain dance rather than, say, doing something materially productive, such as hunting or farming, then the answer seems to be far from trivial.

5.3 Mechanisms in the special sciences

So far we have focused on Cohen's defence of functional explanations against the charge that they are unable to conform to the requirements of the deductive-nomological model of scientific explanations. Cohen proceeds by claiming that functional explanations in the social sciences are analogous to those in evolutionary biology. He argues that since we accept functional explanation in evolutionary biology we should accept them in the social sciences as well.

To see the similarities between the application of a consequence law in the social sciences and in biology, we can compare the explanation of the Hopi's rain dance with an explanation of why giraffes have long necks:

1. IF it is the case that *if* giraffes had long necks, *then* they would be reproductively successful, THEN giraffes have long necks.

2. It is the case that *if* giraffes had long necks, *then* they would be reproductively successful.

3. Therefore, giraffes have long necks.[1]

The explanation is structurally analogous to the explanation of the Hopi's performance of a rain dance. The consequence law, 1, connects the disposition that longer necks cause reproductive success with the existence of longer necks at a later time. Initial condition 2 claims that there is a disposition of long necks to cause reproductive success. Finally, since the explanandum, 3, can be deduced from 1 and 2, we have an explanation of the giraffes' long necks.

[1]This should be seen as an (very poor) explanation-sketch. For a more extensive explanation-sketch, see Cohen [1978, p. 269].

Although the explanation of the Hopi's rain dance is structurally analogous to the explanation of the giraffes' long necks, it has been objected that there is a crucial difference between the two explanations. Elster [1980, p. 125] has argued that the reason why the consequence law expressed by 1 is legitimate, is that we have epistemic access to the mechanism that connects the disposition mentioned by the antecedent with the occurrence of the trait mentioned by the consequent. The mechanism is none other than natural selection connecting a trait-type's contribution to reproductive success with the occurrence of the trait-type at a later time. Elster goes on to claim that "Cohen does not, however, provide any similar mechanism for functional explanation in the social sciences, and therefore his argument cannot succeed." [Elster, 1980, p. 125-6]

This objection is related to a general worry that the DN-model provides an inadequate model for scientific explanations. Sylvain Bromberger [1966] has, for example, pointed out that if we can explain why the shadow of the Empire State Building is x meters long with the help of the position of the sun, the height of the building, and the laws of optics; then the DN-model allows us to reverse the explanation and explain the height of the Empire State Building by citing the length of its shadow, the position of the sun, and the law of optics. It is argued that this example shows that the DN-model fails to account for the asymmetrical character of explanations.

Another objection against the DN-model's is offered by Wesley Salmon [2006, p. 50] in the form of the following example:

1. Nobody who takes birth control pills gets pregnant.

2. Che takes his wife's birth control pills.

3. Therefore, Che does not get pregnant.

Although the explanandum follows from the explanans, it does not seem as the explanans explains why Che does not become pregnant. The relevant explanation of why Che does not get pregnant is not that he takes birth control pills, but rather that Che is a man and men do not get pregnant. Salmon argues that the example shows that the DN-model fails to take explanatory relevance into account.

A third example that shows that the DN-model can lead us astray is the following explanation of rain:

1. Whenever the barometer falls, it will soon rain.

2. The barometer falls.

3. Therefore, it will soon rain.

Although the explanandum is deducible from the true explanans, it does not seem as the fact that the barometer falls explains why it will soon rain. In order for the explanans to properly explain the explanandum, there has to be a causal connection between the two, and even if the fall of the mercury in the barometer is highly correlated with rain, there does not seem to be a causal connection between the two events. Rather, it is the decrease in air pressure that is causally connected to both the fall of the barometer's mercury and the rain. Therefore, it is argued, the appropriate explanation of why the mercury in the barometer falls and why it will soon rain is that the air pressure has fallen. If two events are highly correlated without being causally connected, they are said to be spuriously related to each other. It has been objected that the DN-model's failure to distinguish between spurious correlations and explanatory causal connections is a big problem.

Since the DN-model fails to distinguish between genuine explanations, on the one hand, and necessitations, irrelevant explanations and spurious correlations, on the other, it has been argued that the use of the model has to be completed by a story that shows *how* the explanans leads up to the explanandum. Since there are no direct causal connections from the above explanans to their respective explanandum, this requirement will provide us with a way of ruling out the explanations of the height of the Empire State building in terms of its shadow, the infertility of the man in terms of birth control pills, and the rain in terms of the fall of the barometer.[1] The requirement is sometimes formulated metaphorically as demanding that a proper explanation should specify, so to speak, "the cogs and wheels" of the causal processes through which the outcome to be explained was brought about.[2]

Elster's objection is thus that Cohen has failed to offer an account of the cogs and wheels linking the functional explanans with the explanandum, and therefore he has failed to show that the deduction of the Hopi rain dance from **L** and **C** is explanatory relevant. Elster's objection can, as we will see, be given two interpretations, where each interpretation corresponds to different reasons for why mechanisms are needed.

5.3.1 Mechanisms

Before we turn to Elster's argument, let us have a look at what a mechanism is, and investigate whether Cohen has really failed to provide one. Elster has characterized the explanatory role of a mechanism as follows:

> To explain is to provide a mechanism, to open up the black box and show the nuts and bolts, the cogs and wheels of the internal

[1] For a contrary opinion see van Fraassen [1980, p. 132-5].

[2] See, e.g., Peter Hedström and Petri Ylikoski [2010, p. 51].

> machinery. [...] A mechanism provides a continuous and contiguous chain of causal or intentional links [between the *explanans* and the *explanandum*]. [Elster, 1983a, p. 24]

Although there is something intuitively appealing with the demand that an explanation should "open up the black box" and uncover the underlying "cogs and wheels," it is obviously formulated in terms of metaphors in need of further explication.

Several interpretations of these metaphors can be found in Peter Hedstöm and Petri Ylikoski's [2010, p. 51] literature review on social mechanisms.[1] Let us here, however, focus on Hedström's own definition:

> Mechanisms consist of entities (and their properties) and the activities that these entities engage in, either by themselves or in concert with other entities. These activities bring about change, and the type of change brought about depends on the properties of the entities and how the entities are organised spatially and temporally. [Hedström and Ylikoski, 2010]

It does not seem possible to use this definition to rule out Cohen's account of functional explanations. In order to provide a mechanistic account of a rain dance, we need entities (of some sort) that engage in activities (of some sort) that finally produce a rain dance. With respect to the Hopi, we could say that Hopi society and its members are our entities, and rain dances and increased social cohesion are our changes and activities. One way of organising the entities and activities would be to follow Cohen and assign to Hopi society the disposition that whenever a rain dance is danced social cohesion is increased, and that this disposition causes a new rain dance to be produced at a later time. Speaking metaphorically, we could say that when the disposition-cog, consisting of rain dance- and social cohesion-cogs, turns it causes the rain dance-cog to turn.

[1] For example, the following:

1. A mechanism is a process in a concrete system that is capable of bringing about or preventing some change in the system. [Bunge, 1997]
2. Mechanisms are entities and activities organised such that they produce regular changes from start to finish. [Machamer et al., 2000]
3. A mechanism for a behaviour is a complex system that produces that behaviour by the interaction of several parts, where the interactions between the parts can be characterized by direct, invariant, change-relating generalisations. [Glennan, 2002]

In some sense, it seem as what we have offered is an account that conforms to Elster's definition of a mechanism. After all, we have opened the black box of the social world and showed that the movements of the cogs and wheels of Hopi society at time t_1 causing beneficial effects at t_2, cause the cogs and wheels to turn at t_3 so that they produce a new rain dance. Although it is not obvious that Cohen has offered a mechanism in terms of Elster's intentional links between the explanans and explanandum, it seems as he has offered us a causal link. Furthermore, even if the causal link is not described in terms of individuals, it is a causal link between Hopi society's disposition and the dancing of rain dances.

In order to illustrate the causal mechanism-interpretation of a functional explanations, let us use Harold Kincaid's [1996] characterisation of a functional explanation in terms of causes as a possible mechanistic account of Cohen's consequence law:

1. Rain dancing causes increased social cohesion.

2. Rain dancing persists because it causes social cohesion.

We can call this *the macro-mechanism* that connects Hopi society's disposition with the performance of the rain dance.

It is no secret that Elster, as a methodological individualist, has little patience for explanations in terms of social entities, and that his demand for mechanisms is supposed to rule out explanations in terms of social entities. The same idea seems to lie behind the mechanism account of explanation in the social sciences. Hedström and Ylikoski [2010, p. 59], for example, state that "[a] basic point of the mechanism perspective is that explanations that simply relate macro properties to each other [...] are unsatisfactory."

The latter is a bit surprising if we take into account that Hedström and Ylikoski [2010, p. 52] have been careful to point out that mechanisms form a hierarchy where a mechanism at one level presupposes a lower-level mechanism explaining the workings of the entities involved in it. They do also point out that it is not necessary that the mechanisms an explanation appeals to are themselves explained. The purpose of these remarks seems to be to block a reduction of all sociological mechanisms into psychological mechanisms and from there to biological, to chemical, to physical mechanisms. In order to rule out explanations in terms of social entities while at the same time retaining explanations in terms of individuals, we would need some non-arbitrary way of determining the correct level of mechanistic analysis. We will in section 5.5 see one argument that gives us one, although far from conclusive, reason for preferring mechanisms on lower levels over mechanisms on higher levels. For now, however, it is enough to note that there is no reason to rule out the

functional explanation of the Hopi's rain dance because it fails to provide a mechanism.

5.4 A reductionist dilemma

Let us keep the problem of arbitrariness in mind when we now turn to the first interpretation of the mechanistic objection. A version of the objection is provided by Ann Cudd [2005] who argues that functionalist theories of oppression face the following dilemma:

> either the functionalist theory can be shown to work through the production of a feedback mechanism or it must postulate an emergent social force. [Cudd, 2005, p. 40]

In the case of the Hopi's rain dance, the dilemma requires either that we spell out the micro-mechanism that connects the disposition to the performance of rain dances, or that we postulate the existence of some very strange entity (such as a Hegelian Geist) that produces rain dances.

According to Cudd, the second horn of the dilemma is blocked by "the ontological criterion of causal fundamentalism" stating that "macro-level causes supervene on micro-level ones." This leaves the proponent of a functionalist theory with the alternative of finding the feedback mechanism. The problem, according to Cudd, is that if there is an identifiable feedback mechanism, then it is the mechanism rather than the function that should be pointed to as explanatory. A similar point is made by Elster [1980, p. 126] when he states that he is "at a loss to see why functional explanation [sic!] should be of interest over an above the particular mechanisms that may justify it in any given case." In other words, either functionalism entails a commitment to ontologically strange entities or the function should be reduced to causal mechanisms.

There are two ways out of the dilemma. The first is to show that there are reasons to believe that the ontological criterion of causal fundamentalism is unacceptable. The second is to show that even if the criterion is acceptable, it does not (without further assumptions) rule out functions as explanatory.

Cudd's version claims that

> CAUSAL FUNDAMENTALISM: macro-level causes supervene on micro-level ones. [Cudd, 2005, p. 38]

This claim, as we shall see, is more or less obviously true. The question is, however, what it is capable of ruling out.

Since Cudd does not offer a definition of supervenience, let us assume that she accepts the standard definition: a set of *A*-properties supervenes on

another set of *B*-properties if, and only if, no two things can differ with respect to *A*-properties without also differing with respect to *B*-properties. The *A*-properties are called supervenient properties, and the *B*-properties are called base or subvenient properties. For example, it is sometimes claimed that mental properties supervene on brain-states. This means that, e.g., no two beliefs can differ without there also being some difference in terms of brain states.

Understood this way, causal fundamentalism is meant to rule out all claims involving macro-level causes that exist independently of micro-level causes. For Cudd (as for Elster, Hedström, and Ylikoski), the relevant subvenient micro-level causes in the social sciences are composed of individuals (and their properties). In other words, only causes that supervene on causes that involve individuals and their properties can count as macro-level causes. For example, an institution can count as a macro-level cause only in so far as it supervenes on individuals and their properties. With respect to the Hopi's rain dance, the criterion would rule out the claim that Hopi society exists independently of the Hopi individuals living in it.

Although causal fundamentalism is hard to deny, it does not (as it stand) rule out *explanations* that treat Hopi society as something that exists independently of its members. The only thing it tells is that Hopi society does not exist independently of its members. However, considering that the gunman theory of oppression used false assumptions to explain persisting oppression (as we saw in section 3.5), this cannot by itself be sufficient to rule out functional claims. After all, if the assumption that Hopi society exists as an independent entity allows us to offer interesting explanations and good predictions, while at the same time allowing us a high degree of theoretical unification, then we might be willing to include the false assumptions in our explanations for the same reasons as we included the false assumption that agents are rational into our explanations in chapter 3.

However, when we argued that models using false assumptions could count as explanatory in section 3.5, we compared the explanatory power of different model worlds with respect to their credibility. We said that

> a model world, v, is more credible than another world, v', if, and only if, v is closer to the actual world than v' is.

The more credible a model world is, the higher its explanatory power in the actual world. Since it is plausible to assume that causal fundamentalism holds in the actual world, v, it could perhaps be argued that any world, v', where causal fundamentalism does not hold must be very far away from the actual world. Furthermore, it could be argued that any such world must be further away than any world where other contingent facts are different. For example, worlds where people are fully rational. Thus, it could perhaps be argued that

although we have reason to accept some false claims as explanatory, we cannot accept explanations that assume that Hopi society exists independently of its members.

This might, therefore, justify the following adequacy condition for explanation:

> EXPLANATORY CAUSAL FUNDAMENTALISM: no explanation should include macro-level causes that do not supervene on micro-level causes.

We will accept this criterion as it stands and go on to investigate exactly which explanations it rules out.

5.4.1 Social forces

We have already mentioned that explanatory causal fundamentalism rules out explanations of the Hopi's rain dance which treat Hopi society as an entity existing independently of its members. Cudd, however, uses causal fundamentalism to rule out *all* functionalist theories that have not been shown to work through a feedback mechanism. As representatives of functionalist theories without feedback mechanisms she uses a Hegelian recognition theory of oppression and Foucault's theory of social discipline.

Without going into the details of these two theories, let us point out that although Cudd might be right in assuming that Hegel and Foucault posit macro-causes that do not supervene on micro-causes, it is far from obvious that this is the only interpretation of their theories. About Hegel's recognition theory she says that it

> violates the ontological principle of causal fundamentalism, in that it posits forces at the social level that are emergent from the individual level; that is, *there is no posited causal connection* between the social force of the struggle for recognition and the individuals that compose society. [Cudd, 2005, p. 40, our italics]

Foucault's explanation, in turn, is dismissed as

> there is no agent who is posited to have designed the system for this consequence [of maintaining discipline]; rather the consequence is supposed to explain its own maintenance. Thus, there seems to be some lurking social force involved, yet of indeterminate origin and grain. [Cudd, 2005, p. 40]

Let us begin with the Hegelian recognition theory that is dismissed because there is no posited causal connection between the social force and the

individuals. If Cudd means what is usually meant by supervenience, then it is difficult to see why a lack of posited causal connections violates causal fundamentalism. For example, some of the properties of a statue can be said to supervene on the properties of a lump of clay without any causal relationship holding between the clay-properties and the statue-properties. After all, the clay-properties do not *cause* the statue-properties, rather the clay-properties can be said to *constitute* the statue-properties.

Similarly, a Hegelian may argue that the property of being a Geist supervenes on individualistic properties in the same way as the statue supervenes on the lump of clay. On this interpretation, the Geist and its properties would be constituted by the individuals and their properties. This reading of Hegel could be given some additional support if we consider that Geist can be translated not only as spirit as in ghost, but also as spirit as in team spirit. Although the former seems ontologically queer, the latter is easily treated as supervening on the properties of the team members without there being a causal connection between members and team spirit. Since a supervenience-relation does not entail a causal relation, it is difficult to see why the lack of a posited causal connection between the subvenient and supervenient facts violates explanatory causal fundamentalism.

The dismissal of Foucault highlights another peculiarity in Cudd's use of the criterion of causal fundamentalism. She seems to assume that there are only two possible explanations of why a social system has a function. Either the social system has been designed to perform this function or there exists a social force independently of the subvenient individual properties. This is strange since we know from evolutionary biology that although there is no designer responsible for the evolutionary process, this does not mean that evolution is driven by some mysterious biological force that exists independently of any micro-level causes.

As mentioned in the beginning of this section, Cohen attempts to show that functional explanations in the social sciences work in much the same way as in biology. In order to do this he does not try to provide the exact mechanism. Rather, he argues that it is enough to show that there exists some (perhaps to us unknown) social analogue to natural selection that is neither designed nor independent from micro-causes that connects the disposition with the effect.

Cohen [1978, p. 285] distinguishes between *why-questions* and *how-questions*. Explanations are connected to why-questions of the type "why *P*?", as in "why did the Hopi perform a rain dance?"[1] The answer, "because *Q*", to the why-question tells us that *Q* explains *P*, as in Cohen's answer "because if the disposition (i.e., if rain dances are performed then social cohesion increases) exists,

[1]On explanations as why-questions, see also Hempel [1965], Bromberger [1966], and van Fraassen [1980].

then rain dances are performed, and this disposition exists."

A how-question, on the other hand, asks questions of the type "how does Q explain P?", as in "how does the fact that rain dances increase social cohesion explain the rain dance's occurrence?" When posing a how-question we are, in a sense, interested in dissecting the second nomological arrow of the consequence law:

$$(A \rightarrow B) \rightarrow A.$$

In order to show how the disposition, $A \rightarrow B$, causes A. Cohen calls the answer to a how-question an *elaboration* of the explanation. For example, chance variation and natural selection provides an elaboration of evolutionary explanations in biology. Or to use Cudd's and Elster's vocabulary, providing an elaboration to a functional explanation provides us with the underlying micro-mechanism.

Let us return to Elster's objection that there is a crucial difference between functional explanations in biology and in the social sciences. According to this objection, functional explanations fail because we do not have epistemic access to the analogue of natural selection in the social sciences. Although this might be true, let us for the moment focus on whether the only alternative to design is an emergent force that does not supervene on individuals (or other micro-level causes). Cohen [1978, p. 287-9] offers four elaboration-sketches of functional explanations. Two of these can be used to offer intuitive elaborations of the explanation of the Hopi's rain dance.

According to a *purposive elaboration* of the explanation, the chieftains recognise that there exists a disposition in Hopi society that rain dances lead to increased social cohesion, and they see to it that rain dances are performed in order to increase social cohesion. The second how-explanation is provided by, what Cohen calls, a *Darwinian elaboration*. Here we do not assume that the chieftain, or any other member of Hopi society, purposively try to increase social cohesion. Rather we assume that the Hopi communities exist in a hostile world where a high level of social cohesion is necessary for survival. Thus the environment will select in favour of the communities with high social cohesion, "regardless of the inspiration of the practice." [Cohen, 1978, p. 287-9] Just add the assumption that sometimes chieftains (for one reason or another, or for no reason at all) modify their communities' practices and we will have all the ingredients of a Darwinian elaboration of a functional explanation.

According to both elaborations the social force involved in functionally explaining the Hopi's rain dance obviously supervenes on micro-level causes. In the first elaboration, we are told that the social force supervenes on the beliefs and pro-attitudes of the chieftains. In the second elaboration, it supervenes on the properties of the hostile environment, fitness differences of different rituals, and chance variation. Whether Foucault's and Hegel's explanations can be

offered similar elaborations is an open question. As is the question of whether these elaborations of the Hopi's rain dance are true or mere just-so-stories. The point we want to make here is only that it is possible to provide elaborations where the macro-level causes or social forces that figure in functional explanations supervene on micro-level causes.

However, Cudd does not deny that the social forces involved in functional explanations supervene on micro-level causes. The problem, according to her, is that once we have identified the mechanism, then the functional explanation will be reduced to an explanation in terms of the identified mechanism. According to Cudd, Cohen's elaborations will escape the second horn of the dilemma only by reducing the functional explanation to a feedback mechanism.

5.4.2 Reduction

This brings us to the question of whether the existence of an identifiable feedback mechanism eliminates the explanatory importance of the functional explanation? This depends, it seems, on whether supervenience entails reduction. After all, if the fact that a macro-level cause supervenes on a micro-level cause entails the reduction of the macro to the micro, then the functional explanation of the rain dance will be reduced to its elaboration as soon as we identify the subvenient causes.

The problem, however, is that supervenience is usually not taken to entail reduction. On the contrary, those who wish to deny reduction while retaining the intuition behind a criterion of ontological parsimony often invoke supervenience. Supervenience has, for example, been used in the philosophy of mind to capture the intuition that mental properties are distinct from, but nonetheless nothing over and above physical ones. Closer to what we are interested in, Kincaid [1994] has used supervenience to argue that denying that social properties are reducible to individual properties does not entail a commitment to the thesis that the social consist of anything more than individuals. Since this is how supervenience is usually used it is difficult to see why Cudd claims that the existence of an identifiable mechanism together with causal fundamentalism entails a reduction of the supervenient function to the subvenient mechanism.

If we have a look at Cudd's other desiderata for scientific explanations we will get even more reasons for resisting a reduction of functions to mechanisms. Consider, for example, the following plausible desideratum:

> ONTOLOGICAL PARSIMONY: ontology should be as parsimonious as possible, consistent with the ability to answer the kinds of questions that the theory is devised to answer. [Cudd, 2005, p. 38]

According to ontological parsimony we are entitled to introduce new ontological entities only if they help us answer the kinds of question we are interested in answering. This caveat gives us good reasons for resisting reduction not only in the social sciences, but also in the other sciences as well. After all, the reason why chemistry and biology should not be reduced to physics is that they provide interesting answers and generalisations where physics fails.

In order to see this, let us have a look at how Jerry Fodor [1994] argues for the independence of the special sciences such as biology and the social sciences. He begins by pointing out that the appealing intuition behind reductionism is captured by *token-physicalism*: all events that the sciences talk about are physical events. Although token-physicalism implies that all social events supervene on physical events, it does not imply that all social events are reducible to physical events.

Fodor interprets reductivism as involving the existence of so-called bridge laws that connect the predicates of the higher-order science with the predicates of the lower-order science with the help of biconditionals. The reduction of a social law to a law in physics would have to go through bridge laws that connects the social properties with the physical properties. Let S_1 and S_2 be social properties that are connected by some social law, and let P_1 and P_2 be physical properties. The reduction of the social law through bridge laws would then look as follows:

$$S_1x \leftrightarrow P_1x;$$
$$S_2x \leftrightarrow P_2x;$$
$$P_1x \rightarrow P_2x.$$

Furthermore, Fodor points out that proper natural laws involve only natural kinds. Thus, according to Fodor [1994, p. 689], in order for reductivism to follow, there has to be natural kinds in an ideally completed physics that correspond to each natural kind predicate of any completed social science.

The problem is that natural kinds in the social sciences are *multiple realisable* by physical natural kinds that often have nothing in common. For example, the property "... is money" can be realised by a very large number of different physical properties. So instead of neat bridge laws that connect one social type on the left-hand side with one physical type on the right-hand side, we get a "wild disjunction" of physical types on the right-hand side:

$$S_1x \leftrightarrow P_1x \vee P_2x \vee \cdots \vee P_nx.$$

It seems unlikely that the physical properties on the right-hand side will have anything interesting in common. Furthermore, according to Fodor [1994, p. 691], whether the physical descriptions on the right-hand side have anything in

common is "entirely irrelevant to the truth of the generalisations [made by the social sciences], or to their interestingness, or to their degree of confirmation of, indeed, to any of their epistemologically important properties." He goes on to argue that since the business of the social sciences is to make exactly the kind of interesting generalisations that physics fail to make, reductivism is unacceptable.

Fodor exemplifies the problem with the help of Gresham's law: bad money drives out good. Good money is money where the nominal value is roughly equal to its commodity value, the value of the material it is made. Bad money is money where the commodity value is significantly lower than the nominal value. For example, a clipped gold coin, where a small portion of the gold has been scraped off, is bad money. According to Gresham's law, if the circulating currency consists of both good and bad money, then it will through voluntary monetary exchanges quickly become dominated by bad money. In order to reduce this very simply law to physics, the reductionist would have come up with bridge laws that connect all generalisations in terms of social facts to physical facts. To see the difficulty, take the property of being a monetary exchange. Although it seems reasonable to assume that any event that consists of a monetary exchange has a true description in the vocabulary of physics, it is difficult to see how the reductionist would complete the right-hand side of the following bridge law:

x is a monetary exchange $\leftrightarrow \ldots$.

Note that it is not enough to specify the physical properties of all events that have been monetary exchanges so far, e.g., exchanges involving dollar bills, gold coins, or credit cards. Neither is it enough to specify all events that will ever be monetary exchanges, e.g., exchanges involving cell phones or surgical implanted microchips. Since Gresham's law applies to all *possible* monetary exchanges, the reductionist would have to give a physical description of all possible monetary systems, e.g., exchanges involving strawberry jam-covered shoes and three-legged chickens.

According to Fodor, being wildly disjunctive in this way makes it a bad candidate for being a natural kind in physics and therefore unfit to figure in a physical law. The intuition behind the argument is also seen by considering how much more meaningful the predicate "... is a monetary exchange" is compared to the disjunction in the vocabulary of physics. The whole point of Gresham's law is that monetary exchanges have something interesting in common, and what is interesting about the exchanges is not what they have in common under their physical description. Or to put it in slogan form: reduction to physics is undesirable since *physics* fail to cut the *social* world at its joints.

Recall the caveat of Cudd's desideratum of ontological parsimony stating that the ontology should be as parsimonious as possible while not impeding our ability to answer the kind of questions we are interested in. The criterion seems to be similar to what Fodor has in mind when motivating why we need social kinds. If we return to the issue of whether functions reduce to the underlying feedback mechanisms, the crucial question becomes whether there are any interesting generalisations on the functional level that cannot be captured on the micro-mechanistic level.

In order to see that there are reason to believe that something will be lost in the reduction, let us begin by considering the following scenario:

A. n different mechanisms produce a rain dance in each of n different Hopi villages.

Assume that the occurrence of the rain dance can be subsumed under Cohen's consequence law; or, to use mechanism language, the macro-mechanism described by the consequence law leads up to the rain dance. In scenario A, the macro-mechanism is realised by different micro-mechanism in each village. It might, for instance, be the case that in the first village the macro-mechanism is realised by the micro-mechanism described by a purposive elaboration, in the second village it is realised by the micro-mechanism described by a Darwinian elaboration, and in village three through n by micro-mechanisms m_3 through m_n. The purposive elaboration works through pro-attitudes and beliefs, and the Darwinian elaboration through mutations and selection-mechanisms, but they produce the same result in the two villages. If the elaborations have anything in common, it is that they produce the same result and realise the same function. In other words, it seems as the micro-mechanical descriptions fail to pick out what is interesting about the functional generalisation, namely that the rain dance is performed in both villages because it increases social cohesion.

An additional problem with the reduction of functions can be seen if we consider a second scenario:

B. The purposive elaboration correctly describes the micro-mechanism realising the rain dance in Oraibi, Arizona.

In this scenario, we have one village where one micro-mechanism realises the macro-mechanism. If we are entitled to reducing the functionalist explanation to a micro-mechanism somewhere, it should be here. Reducing the function to the micro-mechanism does, however, involve a problem of truth preservation of counterfactuals.[1] A functionalist theory would support the following counterfactual: "if the chieftain had not understood that rain dances increase social

[1] A similar point is made concerning the reduction of social to individual facts by Kincaid [1994, p. 503].

cohesion, a rain dance would still have been performed." If the function is reduced to a micro-mechanism where a necessary condition for the performance of a rain dance is that the chieftain understands that rain dances increase social cohesion, then it would seem as the reduction does not preserve the truth-value of this counterfactual. Thus, also in the simple case, something that is captured at the functional level is lost in the reduction.

If this is correct, then we should point out that Cohen [1982] concedes too much to Elster when he admits that the possession of a micro-mechanism that connects the explanans with explanandum is a sufficient condition for possessing an explanation. If something is lost in the reduction from the functional macro-level to the micro-level, then the micro-mechanism will fail to capture what is interesting at the higher level. Therefore, knowledge of the micro-mechanism cannot be a sufficient condition for possessing a full explanation of a phenomenon. The same point can be made against Cohen's concession in the introduction to the 2000-edition of *Karl Marx's Theory of History*. There he claims that "[y]ou could be a respectable 'holistic' chemist before chemistry began to expose the structure of the elements, but, once such an advance has been made, to insist on the holistic approach is pure obscurantism." [2000, p. xxvi] In the light of what has been said here, this is only true in so far as the holistic explanations in chemistry fail to give us any additional and interesting information.

These arguments do, of course, depend on whether we accept Cudd's theoretical desiderata. If we believe that there are other reasons for denying the existence of non-individual entities or macro-level causes, then these arguments will leave us unmoved. However, it is only by making the theoretical desiderata explicit that we can determine the correct level of analysis in the mechanistic hierarchy. Once the desiderata have been made explicit we are in a position to evaluate whether they are acceptable, and whether they rule out functional explanations.

Let us recapitulate. Cudd's proposed dilemma proved to be false since it failed to rule out that social forces can supervene on feedback mechanisms without making the social forces redundant. Despite the failure of the dilemma, her theoretical desiderata seemed reasonable and could be used to argue for the exact opposite: functional explanations in general do not commit us to the existence of any ontologically entities over and above individuals and their properties. In the next section, we will see how Elster's desiderata can be used to rule out functional explanations. We will also see that these desiderata are unacceptable.

5.5 An epistemic argument

There is a second way of interpreting Elster's argument that is independent of any ontological considerations. According to this interpretation there are epistemic reasons for dismissing functional explanations unless a micro-mechanism is provided, and this reason would remain even if postulating the existence of social forces were metaphysically unproblematic. It can also be pointed out that this interpretation of Elster's argument does not presuppose a commitment to a mechanistic view of explanations, rather it can be seen as an argument for the use of mechanisms in social scientific explanations.

This interpretation of the argument rests on the assumption that "functional explanations can succeed only if there are reasons for believing in a [mechanism connecting] the consequence to the phenomenon to be explained." [Elster, 1983a, p. 61] This argument is, in turn, justified by Elster's [1985, p. 5] view that the goal of science is to provide laws. To comply with this view, we need to make sure that our lawlike statements really express laws and not mere correlations or necessitation. For example, if some third factor caused both the disposition of the rain dance to produce social cohesion, as well as the rain dance, then the lawlike statement would not be a real law. If this were the case, then although the deduction is valid we would not have a real explanation. Elster argues that we need epistemic access to the mechanisms in order to avoid confusing explanation with spurious correlation or necessitation.

However, as we have already seen, not all mechanisms make Elster's cut. A mechanism will help us avoid confusing explanation with correlation only if it reduces the time-span between the explanans and its explanandum. In the social sciences, the best we can do, according to Elster, is to provide mechanisms in terms of individuals and their properties. Therefore, his objection to the functional explanation of the Hopi rain dance can be restated as follows: unless a credible mechanism in terms of individuals is provided, we are not justified in believing the explanation.

Elster [1980, p. 126] illustrates the difference between his and Cohen's positions with the help of their attitudes towards pre-Darwinian explanations in biology. Elster writes that

> the apparent adaptation of organisms to their environment made most biologists consider that features of the organisms could in fact be explained through this ecological adaptation. Darwin showed us that they were wrong, and that reproductive rather than ecological adaptation is the maximand.

So far Elster and Cohen would probably be in agreement. But Elster goes on to state that the pre-Darwinian biologists

were not only wrong, but also unjustified. [Italics in original.]

If Elster is correct, then the functional explanation of the Hopi's rain dance and all other functional explanations are in trouble. After all, the social sciences have not found the social analogue to natural selection yet.

Let us begin by acknowledging that there is something to the claim that knowledge of the underlying micro-mechanism allows us to reduce the risk of confusing explanation with correlation and necessitation. However, the obvious question that springs to mind is whether risking confusing explanation with correlation is such a bad thing that we can never be justified in believing an explanation unless we have first ruled out this possibility? Consider another of Cudd's plausible theoretical desideratum:

> THEORETICAL UNIFICATION: explanation of one level of phenomena should be consistent with, and mutually inform, levels of phenomena above and below it. [Cudd, 2005, p. 38]

Imagine that we have a well-corroborated lawlike statement that, if correct, would allow us to unify theories from different fields of the social sciences, and thus increase the level of theoretical unification. According to the theoretical unification desideratum this gives us reasons for accepting the lawlike statement as explanatory.

Now imagine that we lack access to the mechanism in terms of individuals and their properties that Elster demands. That we lack a mechanism might give us *one* reason to reject the explanation. But does it really override all other reasons to accept an explanation, including the fact that it would lead to an increased level of unification? What if the same lawlike statement would allow us to explain many previously unexplainable phenomena and make accurate predictions? Requiring a mechanism sounds like a very high price to pay for making sure that our explanations are never confused with correlation.

We find a similar statement in Hedström and Ylikoski [2010, p. 54] where they claim that "the absence of a plausible mechanism linking X to Y gives us a good reasons to be suspicious of the relation being [explanatory]." It seems to be more to the point that the lack of a plausible mechanism gives us *one* reason to doubt that a lawlike statement is explanatory. Although it would be desirable if we were able to identify a function's subvenient micro-mechanism, there might be other theoretical desiderata that can compensate for this shortcoming. For example, a higher degree of theoretical unification or the ability to answer interesting why-questions.

5.6 Summary

In this chapter, we have looked at two arguments against functional explanations in the social sciences. Both arguments were based on the fact that there is a crucial difference between functional explanations in the social sciences and in biology; namely, that the latter has access to the mechanism of natural selection that links the explanans with the explanandum. It was argued that Cohen's consequence law could be spelled out as a mechanism without violating the commonly accepted definitions of mechanisms in the literature. Thus, the objection could not be that functional explanations lack a mechanism. Rather, the objection seems to be that functional explanations do not provide the right kind of mechanism. According to Elster and Cudd, the necessary micro-mechanism must be in terms of individuals and their properties.

The first interpretation of the objection claimed that functional explanations either had to postulate the existence of emergent social forces that do not supervene on micro-level causes, or that the functional explanation was reduced to an explanation in terms of a micro-mechanism. It was shown that, although the argument was based on reasonable premises, the conclusion did not follow. In fact, it was shown that the criteria that were meant to rule out non-reductive functional explanations instead seemed to support them.

The second interpretation of the objection was that without having access to micro-mechanisms we cannot rule out that our lawlike statements are merely describing correlations. We argued that although this might be true, it does not give us an overriding reason to reject all explanations that fail to provide mechanisms; it only shows that it would be desirable to have knowledge of the subvenient micro-mechanism. There may be other reasons for accepting an explanation, such as a higher degree of theoretical unification or the ability to answer interesting why-questions, that would compensate for the lack of knowledge of the underlying micro-mechanism.

All in all, our failure to identify the micro-mechanism, which the functional explanation of the Hopi's rain dance supervenes on, does not give us conclusive reason to dismiss the functional explanation. Whether it is appropriate to explain rain dancing in terms of functions depends on whether it allows us to capture interesting aspects of the social world that explanations on the micro-level fail to capture. This is a question that is independent of whether we have epistemic access to the micro-mechanisms.

However, that there are no problems with functional explanations in the social sciences in general does not necessarily mean that the functional component of the ideology theory of oppression is unproblematic. Part of the reason for why a functional explanation of the Hopi's rain dance was judged appropriate was that we could conceive of some plausible micro-mechanisms that

could realise the function, e.g., a purposive and a Darwinian mechanism. In the case of the functional claim of the ideology theory of oppression, however, it has been argued that there are *no plausible* micro-mechanisms that can realise this function. If this is true, then the ideology theory of oppression will be in serious trouble. In the next chapter, we will investigate whether this claim is true.

6. Ideology demystified

6.1 Introduction

The final argument against the ideology theory of oppression that we will consider has been advanced by Michael Rosen [1996]. Rosen's argument is interesting as it relies on neither reductionism nor the claim that all functional explanations in the social sciences are flawed.

He does not share Elster's concern that we are never justified in accepting an explanation unless we have epistemic access to the underlying mechanism. He does, for example, admit that there is nothing strange about explaining that someone's ability to speak French improved because that person spent some time in France without us having a clue about the physiological mechanism that brought about this improvement [Rosen, 1996, p. 197].

Furthermore, Rosen agrees with Cohen and accepts functional explanations in both biology and the social sciences. He does, for example, accept that many purposive elaborations of functional explanations, such as the elaboration of the functional explanation of the size of car factories from chapter 5, fits neatly into the ordinary *folk psychology* we use to explain everyday actions. Therefore, he accepts that there is nothing strange about functional explanations in the social sciences.

According to Rosen, it is appropriate to functionally explain the size of car factories and the performance of rain dances because it is possible to conceive of reasonable underlying mechanisms that produce these results and that do not entail the existence of ontologically strange entities. The problem with the ideology theory of oppression, for Rosen, is that it entails that people systematically accept beliefs and norms that go against their interests, and that there is no underlying mechanism that can produce this result without having to postulate the existence of ontological queer entities such as Hegelian Geists. He goes on to point out that since there is no place for such entities in our other well-accepted theories, this gives us a strong reason to abandon the ideology theory of oppression.

However, Rosen points out that strange ontological commitments do not provide a conclusive reason to dismiss the ideology theory of oppression. For example, if the only way to explain persistent oppression would be to invoke

ontologically queer entities, then we would be justified in expanding our repertoire of acceptable ontological entities. However, Rosen believes that the gunman theory of oppression can explain everything we are interested in explaining about persistent oppressive social orders; therefore, there is no need to expand our ontology. Rosen concludes that we should give up on the ideology theory of oppression.

In chapter 4, we showed that the gunman theory of oppression cannot explain everything that we are interested in explaining about the persistence of oppressive social orders. Nevertheless, Rosen might argue that being forced to accept the existence of ontologically extravagant entities is very costly in terms of loss of parsimony. The cost may be so high that it cannot be compensated by an increase in other theoretical desiderata. Therefore, we might have no choice but to abandon the ambition of explaining the persistent inequalities in the US and Sweden.

The aim of this chapter is to offer three mechanism-sketches that can realise the function described by the ideology theory of oppression and show that they do not involve ontologically queer entities. Before we move on to the mechanisms, we will in section 6.2 take a closer look at Rosen's argument. Then, in section 6.3, we will propose an evolutionary mechanism based on the theory of the belief-equivalents of genes, so-called *memes*. In section 6.4, we will use concepts from evolutionary game theory to provide a second elaboration of the functional claim. We will show that if the members of a population tend to imitate those who are successful, then the replicator dynamic can be used to describe the growth rate of strategies and ideas in a population. We will then use the replicator dynamic to show how a population can get stuck in an outcome that goes against the oppressed members' overall interests. Finally, in section 6.5, we will show that it is possible to offer a purposive elaboration of the functional claim with the help of the theory of *information cascades*. Although all three mechanisms are able to capture some of the relevant aspects of the ideology theory, the third has the most interesting implications. We will show that it can be used as a micro-foundation for Edward Herman and Noam Chomsky's [1988] so-called *propaganda model* of news media. Section 6.6 concludes.

6.2 Rosen's argument

Rosen's argument against the ideology theory of oppression can be spelled out with three premises and a conclusion:

1. It is possible to explain the persistence of all oppressive social orders without the ideology theory of oppression.

2. By accepting the ideology theory of oppression, we become committed to the existence of ontologically queer entities.

3. We should not accept a theory that commits us to the existence of ontologically queer entities unless this theory is necessary to explain something we are interested in explaining.

4. Therefore, we should not accept the ideology theory of oppression as an explanation of the persistence of oppressive social orders.

In this chapter, we will focus on premise 2. Before we move on to discussing this premise, let us make some brief comments about premises 1 and 3. First of all, we have no quarrel with premise 3 since it is just a restatement of Cudd's very plausible desideratum of ontological parsimony.[1] Premise 1, on the other hand, is not as easy to accept. We have already shown, in chapter 4, that there are good reasons to believe that the ideology theory of oppression is needed to explain interesting aspects of persistent oppression that the gunman theory cannot. We did, for example, find that it is needed to explain the persistent economic inequality in the US, and gender inequality in Sweden. Therefore, in order for the conclusion to follow, Rosen would either have to give us reasons to believe that our conclusions in chapter 4 were wrong, or modify the argument and argue that the benefits of being able to explain every aspect of the persistence of oppression does not compensate for the cost of expanding our ontology if it commits us to the existence of ontologically queer entities. For the sake of the argument, however, we will assume that it would be too costly to accept the existence of such entities. The argument, therefore, hinges on whether the ideology theory of oppression commits us to the existence of ontologically queer entities. Let us, therefore, turn to Rosen's argument for premise 2.

According to Rosen, we become committed to the existence of ontologically queer entities by accepting the functional component of the ideology theory of oppression. To see this, recall that we formulated the ideology theory of oppression in chapter 4 as follows:

> IDEOLOGY THEORY OF OPPRESSION:
> Oppressive societies maintain themselves without depending solely on coercion
>
> 1. in virtue of ideological beliefs and norms among a substantial part of the citizenry, and
> 2. these citizens are subject to ideology because this serves the function of upholding the oppressive status quo.

[1] See page 128.

Consider now the following explication of the functional claim (condition 2) of the ideology theory of oppression with the help of a Cohen-style consequence law:

A. **Consequence law:** IF it is the case that if ideology (at t_1) causes the persistence of the oppressive status quo (at t_2), THEN ideology persists (at t_3).

B. **Disposition:** ideology (at t_1) causes the persistence of the oppressive status quo (at t_2).

C. **Persistence:** Ideology persists (at t_3).

Furthermore, assume that both A and B have gained support from our empirical studies. That is, we have established that 1) whenever ideological beliefs figure among the citizens in an oppressive social order, then the oppressive social order persists (condition B), and 2) whenever this disposition exists in an oppressive social order, then ideology persists over time (condition A).[1] According to Cohen, this would be enough to explain the persistence of both the ideological beliefs and the oppressive social order.

This explanation would be analogous to the following valid explanation of why car factories operate on a large scale:

D. **Factory consequence law:** IF it is the case that if production on a large scale (at t_1) causes persistence of high profits (at t_2), THEN production on a large scale persist (at t_3).

E. **Factory disposition:** Production on a large scale (at t_1) causes persistence of high profits (at t_2).

F. **Persistence:** Production on a large scale persists (at t_3).

If we successfully establish the disposition and the consequence law (conditions D and E), then we would have a proper explanation of why car factories operate on a large scale.

However, Rosen argues that the reason the functional explanation of the scale of car factories is valid, is because Cohen has offered at least some possible elaborations of the nomological connection in D between the disposition and the fact that car factories keep on operating on a large scale. The first in terms of the purposive actions of the managers who recognise the existence of the disposition and act accordingly. The second in terms of an evolutionary mechanism where the disposition together with a hostile environment that

[1] In chapter 7, we will indicate how this can be done using regression analysis.

eliminates unprofitable car factories make sure that only car factories producing on a large scale persist. Although we may not be able to determine whether it is the purposive or the evolutionary mechanism that realise the function in each individual car factory case, we are justified in offering this functional explanation because we know that there is a possible underlying mechanism spelled out in non-mysterious terms.

Rosen goes on to argue that Cohen's suggested elaborations fail in the case of the functional claim of the ideology theory; therefore he concludes that the only way of endorsing the claim is by accepting a mechanism that requires us to accept ontologically queer entities. Rosen's argument rests on the premise that there are only two ontologically harmless elaborations (the purposive and the evolutionary) of the functional claim of the ideology theory. One way of objecting to Rosen would thus be to reject this implicit premise. Here, however, we will show that elaborations of both evolutionary and purposive types are available to a proponent of the ideology theory of oppression.

The main target of Rosen's objection is Cohen's evolutionary elaboration. Cohen motivates the evolutionary elaboration of the Marxist functional explanations by arguing that

> there are traces in Marx of a Darwinian mechanism, a notion that thought-systems are produced in comparative independence from social constraints, but persist and gain social life following a filtration process which selects those well adapted for ideological service. [Cohen, 1978, p. 291]

In other words, the productive forces of a society determine the ideologies that will persist in much the same way as the natural environment determine the traits that will persist in a population. Just as the industrial revolution changed the colour of the trees around the Manchester area, and thus caused a change in colour of the peppered moth population, so did it change the means of production, and thus changed the dominant beliefs and attitudes concerning work and capital.

Rosen [1996, p. 198-200] advances two objections against an evolutionary elaboration of the functional claim of the ideology theory. The first is that when it comes to ideology the time span is too short to allow for the random mutations and filtration processes that are characteristic of the Darwinian struggle for life. Rosen argues that even if we discovered a correlation between bourgeois individualism (ideology) and the development of capitalism (oppression), and hypothesise that the former exists because it maintains the latter, we would not be able to offer a Darwinian elaboration of the explanation since the ideology arose too soon after the development of capitalism. There are at least two responses to Rosen's first objection.

First, there is no reason to assume that the frequency of mutations and reproduction in a population of ideas is as low as in a human population. Even if there would be no time for mutations to occur and go through a filtering process in a human population (or any other animal population), there could be plenty of time for this to happen in a population of ideas. Just as banana flies evolve quicker than humans, and bacteria evolve quicker than banana flies, so might ideas evolve quicker (or slower) than any of these organisms. Think for example of so-called internet memes that seem to evolve at extreme speeds.

Second, although ideas may take a long time to get formed there is nothing in the evolutionary elaboration of the ideology theory that requires that the ideas are formulated in response to the changes in the means of production. The only thing that is required is that they are selected in response to these changes. So if the ideas already exist as part of our common stock of culture then it is enough that they are selected under a short time-span. The filtration process then explains not how ideas are generated, but only how they are accepted and successfully spread in a population.

This second response is anticipated by Rosen and brings him to his second objection against the evolutionary elaboration of the ideology theory. According to Rosen, a crucial feature of the ideology theory of oppression is that ideologies are accepted by those whose interests they go against [1996, p. 198]. He goes on to argue, that it is far from obvious how an evolutionary theory can explain such a 'perverse' phenomenon. Rosen does, however, acknowledge that as a general claim it is false that evolutionary theory cannot explain how a trait, that is obviously bad for an individual animal or for a population of animals, would be selected. The reproductive inability of worker bees is an example of a trait that is disadvantageous to the individual bee yet selected for by evolution. Stephen Jay Gould's [1974] explanation of the antler size of the now extinct Irish elk provides an illustration of how a trait that is bad for species can be selected by evolution. The mating rituals of the elks included jousting to impress the females. Since larger antlers increased the chance of winning the jousts, larger antlers also made their owner more reproductively successful. Therefore, according to Gould, constant sexual selection increased the size of the antlers to the point where they could not go through with the routines of daily life and thus became extinct.

The story about the Irish elks illustrates the point that genes are selfish replicators caring neither about individuals nor populations. Thus, in evolutionary biology the traits that are selected are those that continue to be selected and transmitted, whether or not they are advantageous to an individual or a population. According to Rosen [1996, p. 199], however, an evolutionary explanation of ideology is different from an evolutionary explanation in biology since "the reason why ideas will be selected will be the perceived advantage

that they bring to the thinker." Therefore, he continues, ideas that go against the interest of the thinker must be either the result of a failure of rationality or of some discrepancy between real and perceived interests. While he admits that there are some cases where such discrepancies arise, he argues that this

> is far from showing that (or how) capitalism creates a framework such that individuals *in general* voluntarily and in virtue of the normal mechanisms of belief-formation accept ideas that go against their interests. [Rosen, 1996, p. 199]

Let us ignore the problem of voluntary belief adoption and assume that Rosen's objection is that normal mechanisms of belief formation do not run the errands of capitalist oppressors. He concludes that since we have no reason to believe that the ideology theory of oppression can be combined with any commonly accepted theory of belief formation, accepting it will require an undesirable ontological commitment from us.

Although Rosen captures some of the intuitions behind the suspicion against the ideology theory, he does not treat the possibility of providing an evolutionary elaboration fairly. In the next two sections, we will have a look at two evolutionary elaborations of the ideology theory of oppression that both to some extent escape Rosen's objection.

6.3 Memetics

One possible elaboration of the ideology theory of oppression can be based on the theory of *memes* suggested by Richard Dawkins [2006, ch. 11]. According to Dawkins, evolution by natural selection is *substrate neutral* in the sense that the same explanatory structure can be applied to other domains than biological evolution. He suggests that it can be fruitful to use the same theory to explain, for example, the persistence of cultural traits as we use to explain the persistence of biological traits. In order to illustrate the idea, Dawkins introduces the term "meme" as a belief equivalent of a gene and argues that the prevalence of belief systems in a population can be explained in terms of relative reproductive success of the belief system's associated meme.

We can illustrate memetics with the help of what probably is the memeticists favourite past-time: analysing the spread of religious beliefs. It can, for example, be argued that one of the reasons that the Abrahamic religions are widespread is the first commandment's requirement to not worship other gods. In a sense, this lends the monotheistic belief systems some immunity from invading religion memes that pagan polytheistic belief systems lack. It could also be argued that Christianity's admiration of doctrines such as *certum est*

quia impossibile est (it is certain because it is impossible) also helps it survive invasions by competing memes.[1]

However, one of the best examples of a memetic theory of how religion memes spread is provided by Hugh Pyper's [1998] explanation of the reproductive success of the Bible. He argues that the explanation of the Bible's extreme reproductive success can be traced to its ability to alter the environment in a way that increases its chances of being copied. It does, for example, contain the instructions that it should be passed on, it describes itself as indispensable to its readers, and it is very adaptable so that it can be used to justify almost any action or moral view.

Susan Blackmore [2000, p. 189] identifies yet another "meme-trick" to account for the replicatory success of the major religion memes. Many of the religion memes employ, what she calls, the beauty-trick that inspires the faithful to build impressive buildings and statues in the name of Buddha, Jesus Christ, or Muhammad. The beautiful constructs help spread the memes by inspiring those who lay their eyes on them. According to memeticists, one of the reasons why religion memes are successful is because they have the ability to change their environment to facilitate further proliferation.

Just as genetics takes the gene's eye view in the sense that it focuses on the reproductive success of genes, so does memetics take the meme's eye view and focus on the reproductive success of the memes. Thus, if we can show that an oppressive ideological belief system's associated meme has a higher relative reproductive success than a non-ideological belief system then we will have a memetic explanation of the oppressive ideology even if it is detrimental to the interests of the infected individuals.

Remember that we argued that the widespread belief in the American dream contributed to the explanation of the persistence of the economic inequality in the US. We also argued that the American dream is accepted because it contributes to the survival of the unequal social order. We can now use the memetic framework to provide a sketch of the connection between the American society's disposition to reproduce inequality when its citizens believe the American dream, and the widespread belief in the American dream.

If religion memes alter their environment to facilitate their reproduction, then it is likely that other successful ideas do the same. In this case we could argue that the success of the American dream-meme indicates that it has successfully promotes a social order where its reproductive success exceeds that of competing memes. An unequal social order seems like a place where the belief that anyone can make it through hard work can thrive. For example, in this environment the meme can spread by piggybacking on people's need to

[1] See, e.g., Dawkins [1993].

believe that they are in control of their own lives. The meme can also spread by instructing the few hosts that have escaped poverty to display their wealth and be proud of their poor upbringings. The few who actually make it will become visible and thus reinforce the belief in the American dream. Finally, the meme seems to employ a version of the beauty trick that causes the infected to produce works of fiction based on the theme of the American dream. Upon watching the movies and reading the books, more people will become infected by the belief that the US is a country with a high degree of social mobility.

Memetics is attractive since it offers explanations cast in a terminology familiar from evolutionary biology to a wide range of phenomena; everything from the spread of YouTube-clips to the reproduction of religious practices. Furthermore, since memetics focuses on the meme's "interests," it allows us to explain why people accept ideas that go against their interests. After all, as long as it is able to successfully reproduce itself the American dream-meme does not care about the interests of the people it infects.

It does, however, seem as Rosen would be justified in asking whether memetics can really show how individuals in general voluntarily and in virtue of the normal mechanisms of belief-formation accept ideas that go against their interests. There are, after all, a number of limitations of the gene/meme analogy that casts doubt on whether there really is such a thing as a meme.

If the gene/meme analogy is to hold, then memes, just as genes, must be replicators in the sense that they are units that make copies of themselves. There are, however, reasons to believe that unlike genes, memes are not replicators in this sense. For example, it seems that the imitation of ideas is often too error-prone to underpin replication. Tim Lewens [2012] illustrates the problem with the help of a scenario where he has made a Victoria sponge cake based on his secret family recipe and offers a slice to a friend. Having tasted the cake his friend becomes inspired and attempts to make a Victoria sponge cake of his own. Lewens points out that the chances that his friend will hit upon the exact same recipe will be very low. That ideas, unlike genes, are often slightly or significantly altered when transmitted from one person to another, seems to speak against the gene/meme analogy.

Another problem is that ideas do not always spread because they are literally copied. Dan Sperber [2000] argues that often ideas do not spread because they are copied, but rather because something triggers an idea that already belongs to our culturally shared pattern of thought. For example, the reason why Lewen's friend went home and attempted to make a Victoria sponge cake was not because eating a slice of the cake caused the cake-meme to get copied into his mind. Rather, eating a slice of the cake triggered the friend to use a recipe that already was in his cultural repertoire.

Finally, it has been objected that unlike genes, ideas are rarely copied from

a single source. Lewens [2012] points out that in the realm of biological evolution the evolutionary dynamic often relies on an understanding of genes as discrete, transmittable units. Therefore, if idea-tokens can appear in the mind in virtue of exposure to several sources, then it is unlikely that the spread of ideas will resemble the evolutionary dynamic used in biology.

Even if these problems were successfully addressed, Rosen could ask whether memetics brings anything new to the table. After all, in order for memetics to be of any help it has to get us closer to a satisfactory reply to Rosen's objection. The problem is that even if the memeticists are correct about the existence of memes, they still have to specify how exactly memes are reproduced. "One might worry," as Lewens [2012] puts it, "that memetics merely offers a cosmetic repackaging of a familiar set of stories about cultural change." If this is true, then the memetic framework could in principle be used to formulate both more commonly accepted theories about belief change, such as Bayesian updating, and less kosher theories based on a Hegelian progression. In other words, memetics does not give us reason to favour one over the other.

However, although memetics has not given us a theory of belief change, it has, by suggesting that we should focus on the interests of the meme instead of the interests of the individual host, expanded the set of acceptable theories about belief change. When the individual is the focus, it might seem surprising that agents accept ideas that go against their interests. However, when the focus is shifted to the meme the set of acceptable theories is expanded to include theories of belief formation that takes the memes' "interests" into consideration. In other words, if we take the meme's eye view seriously, then we cannot, as Rosen does, rule out theories of belief formation because they stipulate that people might come to accept beliefs that go against their interests. Whether it is possible to formulate a reasonable mechanism in terms of memes that can realise the function we are interested in is, however, still an open question.

6.4 The replicator dynamic

McKenzie Alexander [2007] has argued that it is possible to make sense of cultural evolution without memes. According to Alexander, cultural evolution should be seen as nothing more than the change of beliefs over time, a phenomenon that can happen for a number of different reasons. People might, for example, i) experiment with new behaviours (the cultural analogue of mutations), ii) deliberately instruct their children (a cultural analogue of reproduction), or iii) consciously imitate other people's behaviour (another analogue of reproduction).[1] Furthermore, he argues that since changes in behaviour corre-

[1] See Alexander [2007, p. 19], the list is not exhaustive.

spond to changes in underlying beliefs it is possible to study cultural evolution by studying the evolution of behaviour. [Alexander, 2007, p. 32] We will return to this claim at the end of this section.

The second evolutionary elaboration we will suggest is an attempt to model cultural change as driven by imitation of successful behaviour. By explaining the change of behaviour in terms of imitation, agents are no longer reduced to mere victims of infectious memes and beliefs. Rather, they are treated as rational in a limited sense where they choose behaviour by following rules of thumb. The model will also make the mechanism (i.e., imitation) through which behaviour spreads, explicit. If we accept that it is possible to infer beliefs from behaviour, then we might be able to construct a model that describes ideological change.

We could say that this model acknowledges that there is something to Rosen's claim that ideas and behaviour that are in the interest of thinkers have a higher probability of being accepted than ideas that go against their interests (all other things being equal). Rosen's argument can be assumed to consist of the following two claims:

1. it is unlikely that a set of beliefs and values are accepted by an individual if there is another set of beliefs and values that are more in the interests of this individual, and

2. the probability that an individual adopts an idea is proportional to the extent it serves her interests.

Let us call these the proportionality claims.

It has been suggested that if people follow a set of behaviour rules resembling the proportionality claims, a model of cultural evolution can be built around the most common dynamic used in evolutionary modelling, the so-called *replicator dynamic*. The dynamic was introduced by Peter Taylor and Leo Jonker [1978] and can be used to mathematically describe the growth rate of a phenotype in an environment as a function of the phenotype's average payoff in the environment relative to the average payoff to other phenotypes in the same environment.

In order to formulate the replicatory dynamic, we will need some mathematical notation. Let n_i be the number of individuals with i-phenotypes in a population, and let $N = \sum n_i$ be the total number of individuals in the population. We can then say that $s = (s_1, \ldots s_n)$ is the *population state*, where $s_i = n_i/N$ is the proportion of individuals with the i-phenotype in the population. Now let $F(s_i \mid s)$ denote the number of offspring to an individual with the i-phenotype (i.e., her payoff) given the population state s, and $F(s \mid s)$ denote the average payoff for the population. If we assume that individuals breed

true (offspring share their parents' phenotype), then we can write the growth rate over time, $\dot{s}_i$, for the frequency of individuals with the i-phenotype in the population, s, as follows:

$$\dot{s}_i = s_i(F(s_i \mid s) - F(s \mid s)). \tag{6.1}$$

Assuming that there are some individuals in the population with the i-phenotype, $s_i > 0$, the growth rate will be positive, $\dot{s}_i > 0$, (a larger share of the population will have the i-phenotype) if the individuals who have the i-phenotype have more offspring than the population average, $F(s_i \mid s) > F(s \mid s)$. It will be negative, $\dot{s}_i < 0$, if individuals with i-phenotype have fewer offspring than the average $F(s_i \mid s) > F(s \mid s)$. If there is no difference in number of offspring, $F(s_i \mid s) = F(s \mid s)$, then the growth rate will be zero. The growth rate will also be zero if there are no individuals in the population who have the i-phenotype, i.e., if $s_i = 0$.

If we switch perspective from biological to cultural evolution it is relatively straightforward to interpret s_i as the proportion of individuals who use the i-strategy in a population. What is slightly more problematic is how the payoff function, $F(\cdot \mid \cdot)$, should be interpreted. In the biological context, the number of offspring gave us an immediate connection between differences in payoffs and changes in growth rates over time. In the cultural context, it cannot be offspring in the literal sense that drives evolution. After all, cultural traits are not only passed inter-generationally from parent to offspring, they are also passed between individuals in the same generation. Furthermore, unlike biological traits, people seem to have some say over which cultural traits they adopt.

Let us assume that the payoff function is given in terms of welfare and that everyone prefers higher to lower welfare. The connection between welfare differences and growth rates of strategies can then be explained by the fact that successful agents tend to be imitated more often than unsuccessful agents. If people are not, or cannot, be fully informed about the consequences of all possible strategies with respect to welfare, then it might be a bad idea to attempt to figure out the strategy that will maximise their own welfare. After all, they might not even know what strategies are available to them. The best thing to do in this type of situation might be to imitate a person with a high welfare. That is, to adopt the strategy used by more successful agents.

Karl Schlag [1998] has shown that the replicator dynamic can be used to describe the growth rate of strategies in a population if people follow so-called *proportional imitation rules*:

1. change behaviour through imitation of others,

2. never imitate an individual that performed worse than you, and

3. imitate an individual that performed better with a probability that is proportional to how much better this individual performed. [Schlag, 1998, p. 131]

The proportional imitation rules allow us to interpret the replicator dynamic in the cultural evolutionary setting. Assume that before an individual is to play a game, she will be given the opportunity to study another player in her population. By doing so she will be given information about the strategy the other player uses and the average payoff the other player has received when playing this strategy. If she follows the proportional imitation rules then, instead of just being given a strategy, she will switch strategy only if she encounters an individual who does better than she does. Furthermore, it is assumed that the better the other strategy is doing compared to her present strategy, the more likely will it be that she switches. This assumption can be justified by the observation that she might be unsure of the exact payoff she will receive if she switches strategy. Therefore, she might be unwilling to switch if the payoff difference is very small. However, as the payoff difference between her current and the sampled strategy increases, the probability that her actual payoff will increase if she switches increases, and consequently the probability that she will imitate the sampled strategy increases.

Since the proportional imitation rules resemble Rosen's proportionality claims, we can move on to the question at hand: can a population of individuals who follow the proportional imitation rules get stuck in an outcome that does not serve their interests? Consider the game in figure 6.1 with the row-strategies Up and Down and column-strategies Left and Right. Outcome

		Column	
		L	R
Row	U	$3,1$	$0,0$
	D	$0,0$	$2,2$

Figure 6.1: Exploitation game.

(U,L) represents a situation where the row player manages to exploit the work of the column player, and the (D,R) outcome represents an equal distribution of the output. For the sake of simplicity, let us assume that if the players fail to coordinate both will get a zero payoff.

Let us now assume that the population consists of two groups: capitalists, who always play the row-role, and workers, who always play the column-role. Let $x \in [0,1]$ and $y \in [0,1]$ be the proportion of capitalists that play U and workers that play L respectively.[1] Let $s = (x,y)$ be the population state that

[1]Since there are only two strategies, we know that the workers who do not play

describes the proportion of capitalists playing U and workers playing L. For example, population state $s = (1,1)$ means that all capitalists play U and all workers play D.

Now assume that from time to time one individual from each population is chosen to play the game in figure 6.1. Before playing, they get the opportunity to assess and revise their own strategy in the way described by the proportional imitation rules. We can imagine that the individual worker compares her present payoff with the payoff of another worker chosen at random. If the other worker did worse than she did, then she will keep on playing her present strategy. If, on the other hand, the other worker did better, she will switch to the other worker's present strategy with a probability proportional to how much better the other worker did. The same goes for the revising capitalist.

Given the payoffs from the game in figure 6.1 and that workers and capitalists follow the proportional imitation rules, it is possible to derive the following differential equations to describe the changes in the population state:[1]

$$\dot{x} = (3y - 2(1-y))x(1-x) \tag{6.2}$$

$$\dot{y} = (x - 2(1-x))y(1-y) \tag{6.3}$$

Equation (6.2) shows how the proportion of capitalists who play U changes given the present proportion of capitalists who play U and the proportion of workers who play L. Similarly, equation (6.3) shows how the proportion of workers who play L changes given the present population state. The so-called solution orbits of the differential equations are shown in figure 6.2. Each point in the diagram represents a population state, $s = (x,y)$. The arrows in the diagram show the direction of the growth rate for each population state. For example, when 40% of the capitalists use the U-strategy and 30% of the workers use the L-strategy, $s = (0.4, 0.3)$, then the growth rate of both strategies is negative. That is, in the next time period less workers will play L and less capitalists will play U. In other words, the population state is moving towards a state where nobody plays U and L.

If Nash equilibrium is the central concept of standard game theory, evolutionary stable state is the central concept of evolutionary game theory. A population state s^* is said to be evolutionary stable if, and only if, it will return to s^* after it has been invaded by a small number of mutant strategies. In the game in figure 6.1 it is straightforward to show that $s_1 = (1,1)$ and $s_2 = (0,0)$

L will play R, thus the proportion of workers who play R is $(1-y)$. Similarly, the proportion of capitalists who play D is $(1-x)$.

[1]For proof, see Schlag [1998]. Similar dynamics can also be derived from other sets of revision protocols. See, e.g., Weibull [1997, p. 186-93].

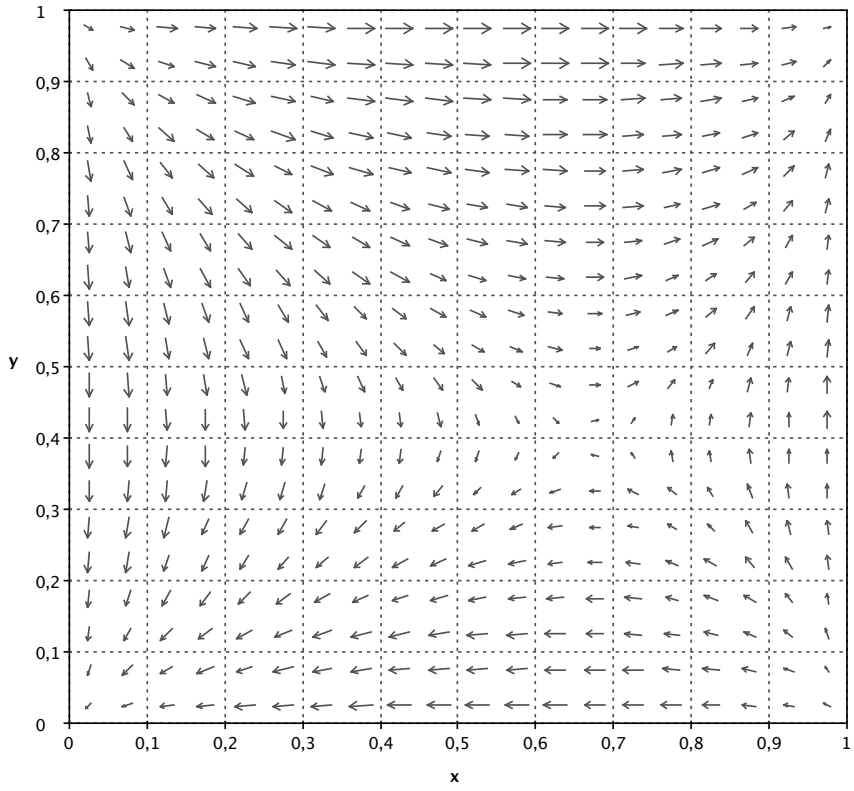

Figure 6.2: Solution orbit for the exploitation game.

are the only two evolutionary stable states.[1] That s_1 and s_2 are evolutionary stable can also be seen in figure 6.2 where $s_1 = (1,1)$ is represented by the upper right corner and $s_2 = (0,0)$ by the lower left corner. That the arrows close to the corners point back towards their respective corner, represent the fact that small changes in the population state will bring it back to the stable state.

What is interesting for our purposes, however, is that the evolutionary model shows that workers can adopt and hold on to a strategy even if it goes against their overall interests. After all, if the choice was between the two evolutionary stable population states s_1 and s_2 the latter would serve the workers interests better than the former, yet they will not switch from s_1 once the population have settled in at s_1. Furthermore, according to the evolutionary dynamic of the game, all population states that are close to s_1 will tend towards s_1. Therefore, the model does not only show that workers who find themselves in s_1 accept L, but also that they will switch to L if they find themselves in a population state close to s_1.

It should also be pointed out that there are good reasons to prefer the model

[1]It is also possible to prove that s is evolutionary stable if, and only if, s is a strict Nash equilibrium. See Jörgen Weibull [1997, p. 167] for a proof.

provided by evolutionary game theory to the model provided by the gunman theory of oppression in chapter 3. Research in evolutionary game theory has to a large extent been motivated by a desire to provide a more realistic foundation for microeconomics than that provided by the assumptions of standard game theory. Instead of the ultra-rational agents interacting under the assumption of common knowledge of standard game theory, evolutionary game theory describes agents as boundedly rational with access to limited information about their opponents and environments.[1] Since there seems to be a lot that speaks in favour of people being more like boundedly rational imitators than ultra-rational maximisers, it could be argued that models provided by evolutionary game theory describe more credible worlds than the models provided by standard game theory.

The problem is that even if we successfully convinced Rosen that the above model is an accurate representation of an oppressive situation, the proportional imitation rules are supposed to result in change of behaviour and not in adoption of ideas. However, as we mentioned in the beginning of this section, Alexander [2007, p. 32] argues that we can use the replicator dynamic to model cultural evolution. He suggests that the connection between behaviour and belief changes is provided by the standard assumptions of folk psychology. Beliefs and pro-attitudes explain action, and if we observe a change in behaviour, then there must be some change in either beliefs or pro-attitudes to explain this. If the pro-attitudes are unchanged, then the change in behaviour can only be explained by a belief change.

In the case of the ideology theory of oppression, this can perhaps be illustrated by the tendency of people to not only imitate the behaviour of successful individuals, but also attempt to emulate their mindset. There is after all a huge market for biographies and self-help books written by, and about, successful people, that explain not only what to do, but also how to think, in order to achieve success. Since the ideas of the rich and famous often have an individualistic streak, it could perhaps be argued that the ideological belief of bourgeoisie individualism spread through imitation of the successful.

If we add the assumption that workers can become capitalists, then we could perhaps argue that the adoption of the capitalists' ideas improve the probability of becoming a capitalist. After all, seeing the world for what it is may cause a worker to become either embittered or politically radicalised. Neither is very attractive in the eyes of employers. A worker who has adopted the capitalists' mindset will, on the other hand, be less hostile and more adaptable, and therefore a more attractive employee. Although all employees will not become CEOs, some could. Since having the right attitude in this type of

[1] See, e.g., Weibull [1998].

social order will increase the worker's chances of becoming a capitalist, she will be rewarded for becoming a victim of ideology. However, if we assume that the probability of becoming a capitalist is very small even after the right mindset has been adopted, then she would still be better off if the social order was changed.[1]

If the evolutionary game theoretic model can explain not only changes in behaviour, but also changes in beliefs and ideas, then it can provide us with a Darwinian elaboration of the functional claim: ideology persists because it allows the oppressive social order to survive. To see this, assume that a population is in the unequal state described by the population state $s = (1,1)$ in the game in figure 6.1. They are stuck in this state in part because the workers are victims of the ideology associated with strategy L. Furthermore, the disposition of the unequal social order to persist when the workers believe the ideology associated with L, explains that workers continue to be victims of ideology. If the workers follow the proportional imitation rules, they will continue to imitate the most successful workers who, given the population state $s = (1,1)$, will continue to be the ones who hold the ideas associated with L. After all, the best response in the neighbourhood of $s = (1,1)$ will be to imitate those who do L. Consequently, any R-invaders, in the form of revolutionary immigrants or radicalised youth, will quickly adapt and do L.

However, the problem for an elaboration of the ideology theory of oppression is that these beliefs are in a sense underdetermined by the behaviour. After all, the very same behaviour can be explained by a great number of different beliefs, and not all of these beliefs are ideological. The problem can be illustrated by a simple example. Assume that we have a model, where the agents are described as following the proportional imitation rules, that can be used to accurately predict the social orders where oppression will persist and where it will disappear. Having access to a mechanism that allows us to make successful predictions seems to speak in favour of the corresponding higher order functional explanation.

The problem is that this functional explanation will not necessarily be an explanation in terms of *ideological beliefs and norms*. The underlying evolutionary model provides us only with a functional explanation in terms of behaviour: political inactivity causes the status quo to persist, and political inactivity persists because it causes the status quo to persist. If we want to use

[1] Assume that if the right mindset is adopted, the probability of becoming a capitalist is γ. Assume also that there is only upward social mobility. A worker who has adopted the right mindset will then get the expected payoff $3\gamma + (1-\gamma)$ in the exploitative social order. In an egalitarian society, on the other hand, she would get a payoff of 2 for sure. Thus only if $\gamma > 0.5$, would the exploitative social order be better for her than the egalitarian social order.

folk psychology to ascribe beliefs to the citizens we do not seem to have any reason to claim that the inactivity was caused by ideological beliefs rather than, e.g., prudential considerations. Rosen could, therefore, argue that nothing has made it more probable that the beliefs causing the behavioural change were ideological.

In other words, the model does not show that oppression persists due to ideological beliefs and norms.[1] The behaviour could as well have been caused by the non-ideological belief that L was the welfare-maximising action given the population state s_1 and the desire to maximise welfare. That is, the model fails to establish how oppressive societies maintain themselves in virtue of ideological beliefs and values.

Rosen [1996, pp. 260-2] suggests that the best way to explain persistent oppressive social orders is in terms of fully rational agents motivated by non-ideological beliefs and prudent desires. He might, therefore, argue that if the evolutionary model can be used to elaborate a functional explanation of persistent oppression, then this will be a functional explanation that supports his case rather than the ideology theory. The model shows how somewhat rational and informed agents can become more and more entrenched in an oppressive social order, it does not show that they are victims of false consciousness.[2]

However, since there are aspects of ideology involved in many persistent oppressive social orders, this conclusion would not be entirely satisfactory. After all, we showed in chapter 4 that rational choice models were unable to explain the persistent economic inequality in the US and gender inequality in Sweden. We should, therefore, press on an attempt to find an elaboration that can show not only how behaviour compatible with ideology can come about, but also how ideological beliefs and norms are produced.

6.5 Ideology production as information cascades

Rosen's preferred alternative to the ideology theory of oppression is the gunman theory of oppression. According to the gunman theory, revolutions fail to materialise because rational agents do not participate in revolutionary action. They abstain because they are subject to free-rider problems. One of the explanation's underlying assumptions is that the oppressed are rational in the sense that they choose the actions that maximise their own expected welfare. It is also worth pointing out that the explanation provided by the gunman theory is

[1] See Bicchieri [2006, p. 217] for a similar objection against the use of evolutionary dynamics to explain fairness norms.

[2] See also Olsson-Yaouzis [2012] for an explanation of the lack of revolutions in tyrannies in terms of a replicator dynamic.

of the same type as the purposive elaboration of the functional explanation of the Hopi's rain dance and the size of the car factories. According to the purposive elaboration of the Hopi's rain dance, the chieftain recognised the existence of the disposition of Hopi society to increase social cohesion whenever a rain dance was performed, and having a desire to maximise social cohesion the chieftain made sure that rain dances were performed.

Rosen [1996, p. 197] barely considers the possibility of extending the purposive elaboration to the functional claim of the ideology theory of oppression. Instead of investigating whether it is possible to provide a purposive elaboration in terms of individuals (as in the case of Hopi's rain dance), he considers the possibility of expanding the purposive framework to collective agents. That is, he considers whether it is possible to ascribe purposes in terms of beliefs and pro-attitudes to groups and societies. He dismisses this possibility with the motivation that we do not have a commonly accepted "folk sociology" that matches the folk psychology we use to explain people's beliefs and actions. He concludes that the only way of explaining social phenomena in terms of the beliefs and pro-attitudes of societies would require us to accept ontologically queer entities.

It is surprising that Rosen does not consider whether the classical rational choice framework focusing on rational individuals can be used to provide an elaboration of the ideology theory of oppression. If we could find an elaboration in terms of rational choice theory, then this would probably be as close as we could get to showing that "individuals in general voluntarily and in virtue of the normal mechanisms of belief-formation accept ideas that go against their interests." [Rosen, 1996, p. 199] After all, if by "normal mechanisms" Rosen means well-accepted mechanisms, then mechanisms cannot get more normal than those offered by rational choice theory. They have been used by economists and sociologists during the better part of the 20^{th} century to explain everything from why drug dealers live with their mothers, to marriage choices, to why taxi drivers are reluctant to pick up members of certain ethnic groups.[1] Furthermore, since rational choice theory provides explicit mechanisms spelled out in familiar terms, e.g., individuals with beliefs and pro-attitudes, a successful elaboration in these terms would allow us to definitely avoid becoming committed to ontologically queer entities.

Let us, therefore, investigate whether it is possible to provide a purposive elaboration within the rational choice framework to the ideology theory of oppression. At first sight, this might seem as a hopeless project. What beliefs and pro-attitudes could explain that the members of the oppressed group rationally

[1]Concerning drug dealers and their moms, see Steven Levitt and Stephen Dubner [2005, ch. 3]. On family issues and rational choice, see Gary Becker [1981]. And finally, on taxi drivers, see Glenn Loury [2002].

come to adopt beliefs that help preserve a social order that is detrimental to their welfare? Unlike the manager who increases the scale of production in the car factory in order to maximise profits, there does not seem to be anything to gain from adopting ideological beliefs. Furthermore, there seems to be a fundamental difference between deciding to believe something and deciding to increase the scale of production. The manager seems to be able to perform the latter at will whereas it is unlikely that an American can come to believe in the American dream at will.

One way of getting around these problems is offered by the theory of so-called *information cascades*. The theory is designed to explain how rational agents come to adopt false or arbitrary beliefs. It has, for example, been used to explain medical fads where certain surgical procedures suddenly increase in popularity, bandwagon effects in US presidential nomination campaigns, peer influence, and bubbles in the financial market.[1]

The basic idea is that if an individual is uncertain about what to do, then the best thing she can do is to observe what others do in similar situations. Observing what someone else does allows her to infer the observed person's belief about the best course of action. She can then use this information to update her own belief about what the best course of action is. We can illustrate this with Che who is unsure of whether it will rain, and therefore cannot decide whether he should bring his umbrella or leave it at home. One way for Che to get additional information is to go to the window and see whether others have brought their umbrellas. Assume that Che sees Fidel walking down the road with an umbrella under his arm. This will allow him to infer that Fidel believes that it will rain. Having received an additional piece of information, Che becomes convinced that it will rain, and therefore decides to bring his umbrella.

The problem is that if everyone has some belief about what to do, but is uncertain about the quality of their own information, then it can be shown that everyone stops taking their own private information into consideration when they make their decision. This can result in situations where everyone follows the herd. Assume, for example, that Raúl believes that it will not rain, but that he is not certain. Just as Che, he goes to the window and there he sees Fidel walk past with an umbrella under his arm. He then infers that Fidel believes that it will rain. Since he has no reason to believe that his own information is better or worse than Fidel's information, he concludes that it is equally probable that it will rain and not rain. At the moment when he decides to let a coin flip decide whether he should bring his umbrella, he sees Che with an umbrella under his arm. Just as he had no reason to believe that his

[1] See, e.g., the two seminal texts on information cascades by Banerjee [1992] and Bikhchandani et al. [1992].

information was better than Fidel's, he has no information to believe that his information is better than Che's. Therefore, the observation of Che is pivotal with respect to Raúl's belief that it will rain. Having inferred that two other individuals believe that it will rain is enough for Raúl to disregard his own initial belief that it will not rain, and he brings his umbrella.[1]

Once rational agents cease to take their private information into account, a *information cascade* occurs. Everyone living down the street from Raúl will end up bringing their umbrella no matter what they initially believed. It is also worth pointing out that the information cascade would have started even if Fidel merely carried an umbrella in order to return it to a friend.

It has been argued that the similar phenomena are responsible for financial bubbles and election outcomes. Since people are unsure of where to invest their money they will try to gather additional information by observing where others invest their money. The problem is that when enough people have made an investment, people will continue to make the same investment no matter their initial beliefs or the actual quality of the investment.

For our purposes, this is a promising theory since it shows that under certain circumstances rational agents adopt false or arbitrary beliefs. Furthermore, although the beliefs are not adopted at will, they can be seen as being adopted as by-products of other purposive processes.

Cristina Bicchieri and Yoshitaka Fukui [1999] have used the theory of information cascades to show how norms can come about and persist in a group although a majority of its members oppose them. Although they are mainly interested in binge drinking among students and bribing in kleptocracies, it seems as their model can be expanded to handle also the cases we are interested in.

The model assumes that the members of a group face a binary choice between x_1 and x_2.[2] Each member is assumed to prefer x_1 to x_2, or x_2 to x_1. Furthermore, the group consists of a large group of *conformists* and a smaller group of *trendsetters*. The conformists always prefer to do what the majority of the members prefers to do. If they are unsure about what the majority wants to do, then they will let a coin flip decide for them. Trendsetters, on the other hand, also want to do what the majority wants to do, but if they are unsure

[1]It is important to note that the reason why it was enough for Raúl to make two observations in order to disregard his initial belief was that he believed that the quality of his own, Che's and Fidel's information was equal. If Raúl is a meteorologist, and therefore believes that the quality of his information is superior to the information of others, then the observation of two umbrellas might not be enough to make him sway.

[2]The formal models of information cascades are often complicated and tend to obscure the underlying intuitions. We will, therefore, settle with an informal version of the model.

about what the majority wants, then they will act on their true preference. It is also assumed that decisions can be made at either t_1 or t_2, where t_1 precedes t_2.

Bicchieri and Fukui assume that the trendsetters have a preference for expressing themselves. Consequently, the trendsetters will choose to act at t_1. The conformists, on the other hand, have no interest in expressing their own preferences. If they wait until t_2 they will get the chance of observing what others do, and thus to increase the probability of doing what the majority wants to do.

Since the trendsetters have no information about what the majority will do (over and above their own preference), they will act according to their true preference. At t_2, the conformists will observe the actions of the trendsetters and then decide what to do. They will use the observations of how the trendsetters acted at t_1 and the knowledge of their own preference to attempt to figure out what the majority wants. If the number of x_1-observations exceeds the number x_2-observations, then the conformist will decide to do x_1. If there are an equal number of x_1 and x_2-observations, then the conformist will let a coin flip decide. Note that the conformist's set of observations contains both what the trendsetters did at t_1 and her own preference. Therefore, if a conformist who privately prefers x_1 to x_2 has observed that four trendsetters did x_1 and that four trendsetters did x_2, then the conformist will form the belief that the majority prefers x_1 to x_2.

Now, let z_i be the number of trendsetters who privately prefer x_i, and therefore do x_i at t_1. It is then relatively easy to show that if $z_1 - z_2 \geq 2$, then all conformists will disregard their own privately held preference and decide to do x_1. The reason is simply that they will have more x_1-observations than x_2-observations even if they privately prefer x_2 to x_1. No matter what they privately prefer, they will go with x_1. In other words, if $z_1 - z_2 \geq 2$ an information cascade will occur. For the same reason, if $z_2 - z_1 \geq 2$, all conformists will decide to do x_2.

If $|z_1 - z_2| \leq 1$, then the conformists will take their privately held preferences into account as part of the evidence for what the majority prefers. If an equal number of trendsetters have chosen x_1 and x_2, then the conformists will act on their privately held preferences. If $z_1 - z_2 = 1$, then the conformists who prefer x_1 to x_2 will decide to do x_1. The conformists who privately prefer x_2, on the other hand, will believe that the outcomes where the majority prefers x_1 and where the majority prefers x_2 are equally probable. They will therefore let a coin flip decide.

For our purposes, the important point is that the conformists will follow the trendsetters no matter what the conformists privately prefer if the number of trendsetters who prefer one alternative to the another is sufficiently large. Also note that although Bicchieri and Fukui's model assumes that the conformists

have an interest in doing whatever the majority prefers, it is relatively easy to extend the model to account for other preferences. The conformists might, for example, have an interest in making the correct investment, vote on the best presidential candidate, or bring an umbrella if, and only if, it will rain. As long as the conformists believe that the trendsetters' private information is at least as good as their own, and that the trendsetters truthfully act on their private information, then rational conformists will become influenced by the trendsetters' decisions.

The assumption that there are some who have a preference for setting trends, and some who have a preference for following them, does not imply that conformists are docile sheep and trendsetter instigators. The preference for following trendsetters can represent the fact that conformists have better things to do than to gather information. It might, therefore, be rational for them to wait and see what the trendsetters do in order to make an informed decision. Similarly, trendsetters' decision to go first does not have to represent a flamboyant desire to manifest their own preferences. A trendsetter may have other incentives to decide early. She may, for example, be an expert in her field, such as a meteorologist or a financial advisor, who will only get paid if she gives early advice, she may be a community leader whose reelection is dependent on whether the community sees her as someone fit to lead, or she may be a peddler in opinions like a lobbyist or a political editor.

If Bicchieri and Fukui are correct about how unpopular norms can spread and persist, then it would be relatively straightforward to extend the model to cover ideological beliefs that maintain oppressive social orders. This would allow us to show how ideological beliefs can become accepted and be transmitted through rational belief updating.

Although the model shows how ideological beliefs and norms can be accepted by rational agents, it does not provide a connection between the persistence of the oppressive social order and the beliefs that will be adopted. Remember that in order to provide a full elaboration of the ideology theory of oppression we would need to show that ideological beliefs persist because they contribute to the survival of the oppressive social order.

We can illustrate the problem with the help of a set of trendsetters who are community leader in an oppressive society. Assume that most citizens initially, with some degree of uncertainty, believe that the regime is unjustified. Since the community members have to take care of their jobs and families they decide to suspend belief until the community leaders have made up their minds. The community leaders are elected in order to provide correct answers and good opinions in this type of situation. Since they want to be re-elected they will gather information to decide whether the regime is justified and then announce their decision to their communities.

In order for the model to provide a complete elaboration of the ideology theory of oppression, it must be able to show how the community members come to believe that the regime is justified. The problem is that if the community leaders consist of a representative sample of the citizens, then most of them will (probably) reach the conclusion that the regime is unjustified.[1] Consequently, contrary to the ideology theory of oppression, the citizens' belief that the regime is unjustified will be reinforced.

However, assume that the leaders of the regime have managed to identify the community leaders and offered them (positive or negative) incentives to declare that the regime is justified. The community leaders will then cease to be a representative sample of the citizens. The citizens who follow their community leaders will thus end up supporting the regime. In other words, it seems as if we manage to identify a (purposive) connection between the oppressive social orders and the trendsetters, then we will be able to provide an elaboration of the crucial functional claim of the ideology theory of oppression.

Fortunately, there is a theory provides a general connection between the representatives of the oppressive social orders and the trendsetters. The so-called *propaganda model* formulated by Edward Herman and Noam Chomsky [1988] seems to be tailor-made for this purpose. They want to explain why mass media reporting tends to favour government and corporate interests. They identify a number of influences on the editors' choices of which news pieces to include and which angles to push for. For example, news media have to deal with i) the political and economic interests of their owners, ii) lobby groups that organise systematic replies to reporters, and iii) the demands of advertisers. [Herman and Chomsky, 1988, p. 2] According to Chomsky and Herman, owners, advertisers and lobby groups tend to side with government and corporate interests and the editors are thus influenced to present the government and corporate friendly interpretation of facts and events.

Although Herman and Chomsky offer a plausible case for editorial bias, we cannot use their model by itself to offer an elaboration of the functional claim. After all, as the propaganda model stands it only states that capitalism creates a framework where news media runs corporate and governmental errands. What Rosen is after is an elaboration that shows how individuals in virtue of normal mechanisms of belief-formation come to accept ideas that go against their interests. Since the propaganda model does not provide an explanation for how the bias spreads from the news room to the citizens, it will fail

[1]Even if most of the citizens believe that the regime is unjustified, there is still a statistical chance that most of community leaders believe that it was justified. [Bicchieri and Fukui, 1999, p. 144]. However, if the number of citizens is very large (i.e., infinitely large), then the proportion of community leaders who believe that the regime is unjustified will be equal to the proportion of citizens who hold this belief.

to convince Rosen.

However, since news media (if anyone) play the role of trendsetters in Western societies we could combine it with Bicchieri and Fukui model to complete the picture and get a purposive elaboration of the ideology theory of oppression. On this interpretation citizens are conformists in the same sense as the community members in the example above. They are therefore reluctant to make up their mind about what to do and think before they have observed what the trendsetters do and say. The editors who want to maximise profits need to both provide biased news and be trendsetters. The representatives of the oppressive social order, in their turn, can be expected to prefer reporting that preserves the social order. Therefore, the norms and ideas that are adopted by the citizens, are the same norms and ideas that preserve the oppressive status quo.

It should also be pointed out that the functional claim that news reporting is biased because it has beneficial effects for the regime can be interpreted evolutionary as well as purposively. The phenomenon that news media tend to be biased towards corporate and governmental interests can be viewed in much the same way as the phenomenon that car factories tend to operate on a large scale. So instead of offering an elaboration in terms of the editors' rational decisions, we could offer an elaboration in terms of variation and survival. On this interpretation, the forces that Herman and Chomsky identify do not constitute reasons for action, but rather they determine the evolutionary fitness of different newspapers. The reason why most newspapers provide biased reporting is, according to the evolutionary elaboration, that only biased newspapers survive in the long run.

News media are, of course, not the only trendsetters available to a regime. Teachers are another group of trendsetters that are under similar pressures as news editors in the sense that if they do not teach according to the desired curriculum they risk losing their jobs. Other groups of trendsetters that regimes might be eager to control are community and religious leaders, authors, filmmakers, etc. Furthermore, if we shift focus from trendsetters in Western democracies, to trendsetters in more tyrannical regimes, we can add threats of imprisonment, torture, and death to the list of forces that influence editors. Consider, for example, the threat of violence journalist in Russia, Eritrea, and Turkey operate under.[1]

Finally, although it might be tempting to believe that the harsher the threat of sanction the more control the regime will have over trendsetters, this is probably not true. The harsher the threats the easier it might become to get

[1]It is worth mentioning, as an ironical anecdote, that the publisher and translators of Herman and Chomsky's *Manufacturing Consent* were in 2006 persecuted (although later acquitted) for insulting Turkishness.

newspapers to publish the desired stories and angles. However, if the harshness of the threats become known, it seems plausible to assume that the journalists become less credible. After all, the readers will then become aware of the fact that the articles are produced in order avoid punishment, and not to convey the author's privately held beliefs and opinions.

A lot more needs to be said about this elaboration. It would, for example, be desirable if the model could be formalised to make the implications clearer. For our purposes, however, it is enough to have shown that it is possible to offer an elaboration of the ideology theory of oppression within the familiar rational choice framework. The development of the formal model will be left for future research for social scientists.

6.6 Summary

In this chapter, we have offered three elaborations of the functional claim of the ideology theory of oppression. None of the elaborations referred to ontologically queer entities. The first mechanism was spelled out in terms of memes and drew attention to the importance of focusing on the meme's eye view when explaining belief change. That is, we should focus on the reproductive success of the belief (or its associated meme) instead of the interest of the believer when explaining belief change. One problem with the memetic approach was, however, that if there were people who found Geists mysterious, then there is no guarantee that they will accept the existence of memes.

The second mechanism showed that we do not need memes in order to explain changes in beliefs in a population. By assuming that people follow so-called proportional imitation rules, we showed how the replicator dynamic could be used to describe changes in behaviour over time in a population. It was then argued that we could use folk psychology to infer belief changes from behaviour changes. Since we could show that people could get locked into an oppressive situations, we could infer that they could also get stuck in ideological mindsets. The problem with this approach was that since each behaviour was compatible with many different beliefs, and that not all of these beliefs are ideological, we could not conclude that the beliefs were ideological.

We did, therefore, move on to a purposive elaboration of the ideology theory of oppression based on the theory of information cascades. Although the third elaboration was far from complete, it provided a way of defending the ideology theory of oppression against Rosen's objection that there was no way to show that "individuals in general voluntarily and in virtue of the normal mechanisms of belief-formation accept ideas that go against their interests." After all, explanations in terms of rational agents may have 99 problems, but invoking ontologically queer entities ain't one.

This chapter concludes our investigation of the methodological problems of the ideology theory of oppression. Having provided three possible elaborations of the theory, we have shown that it is indeed possible to accept the ideology theory without becoming committed to ontological queer entities. Although there were problems with two of them, these problems were not as severe as Rosen seems to believe. What is important is that we have shown that the functional claim of the ideology theory of oppression requires no stranger commitments then other functional explanations in the social sciences and biology.

In the concluding chapter, we will discuss the implications of having vindicated the ideology theory of oppression. We will begin by providing a statistical method that allows us to investigate i) whether ideological beliefs and norms exist in a social order, and ii) whether they exist because they contribute to the survival of the social order. We will end the book by briefly investigating what ought to be done if we discover that oppressive social orders persist through ideological beliefs and norms.

7. What is to be done?

7.1 Introduction

We are approaching the end of our investigation of the gunman and ideology theories of oppression. All that is left is to sum up the arguments and discuss potential implications of the ideology theory of oppression.

We will, in this chapter, begin by providing a brief summary of the arguments in the thesis. We will then present and answer a practical objection against the ideology theory of oppression. We will conclude the chapter, and the book, by discussing what ought to be done if the ideology theory proves to be correct.

Let us begin by recalling that we, in chapter 1, formulated the argument against the ideology theory as follows:

1. It is possible to explain the persistence of all oppressive social orders without the ideology theory of oppression.

2. By accepting the ideology theory of oppression, we become committed to the existence of ontologically queer entities.

3. We should not accept a theory that commits us to the existence of ontologically queer entities unless this is necessary in order to explain something we are interested in explaining.

4. Therefore, we should not accept the ideology theory of oppression as an explanation of the persistence of oppressive social orders.

We began our investigation in chapter 2 by introducing three cases of persistent oppressive social orders that our theories should be able to explain: tyranny in North Korea, economic inequality in the US, and gender inequality in Sweden. In chapter 3, we discussed the gunman theory of oppression at some length and showed that it was ideal for explaining the persistent tyranny in North Korea. We then proceeded to chapter 4 where we showed that the gunman theory is unable to explain the other two cases of persistent oppression. We also showed that the ideology theory of oppression is able to explain both the persistent economic inequality in the US and the persistent gender inequality in Sweden. We concluded that premise 1 was false.

In chapter 6, we showed that the explanation provided by the ideology theory of oppression is neither better nor worse than other functional explanations in the social sciences when it comes to assuming ontological strange entities. This would have been a hollow victory unless we had already, in chapter 5, provided reasons for accepting functional explanations in the social sciences. Having shown that premises 1 and 2 are false, we can dismiss conclusion 4 while retaining the plausible methodological principle expressed by 3.

However, having shown that the ideology theory of oppression does not commit us to the existence of ontologically queer entities does not provide any reason for believing that the theory is actually true. After all, a theory that assumes that cows attempt to maximise their consumption of strawberry jam does not imply any commitment to queer ontological entities. However, this fact alone does not give us any reason to believe that the theory is true. It would indeed be falsified as soon as it was subjected to empirical tests.

In order to test the ideology theory of oppression, we would similarly have to subject it to empirical tests. Now, we could happily hand over this to the social sciences since this is a dissertation in philosophy, and this type of empirical investigations are better conducted by social scientists than philosophers. However, it might be objected that even if there is nothing wrong in principle with the ideology theory of oppression, it is in practice impossible to test the theory and it should therefore be dismissed for all practical purposes. Since this is a conclusion we want to avoid, we will in section 7.2 indicate how the theory could be tested.

We will conclude the chapter, and our investigation, by discussing some of the potential implications for public policy if oppressive social orders persist because a substantial part of the citizens are victims of ideology. In section 7.3, we will discuss what a group of revolutionaries should do if they discover that ideology keeps their fellow oppressed from participating in revolutionary action. We will see that there are reasons to believe that the only way of exchanging an oppressive social order for a less oppressive, is to act against the expressed will of the ideology-ridden oppressed.

7.2 Confirming the ideology theory of oppression

The fact that the ideology theory of oppression does not commit us to a strange ontology should not by itself enough to convince us that the theory is true. In order to become convinced, we would need evidence that indicates that 1) some oppressive social orders persist because the oppressed are victims of ideology, and 2) they are victims of ideology because this maintains the oppressive social order.

In order to establish the truth of these conditions, we would have to appeal

to empirical facts. Although we will not conduct an empirical investigation, we will show that there are no problems associated with testing the ideology theory of oppression that set it apart from other theories in the social sciences.

Kincaid [1996, p. 115] suggests that functional claims, just as any other empirical claim, can be vindicated by evidence. The evidence can be in the form of *indirect evidence* that shows that the functional claim is compatible with another set of facts, or in the form of *direct evidence* that shows that the conditions of the functional explanation hold one by one. Most of the time, however, the evidence falls somewhere on a continuum between direct and indirect evidence.

We can provide indirect evidence for a theory by showing that it can account for our observations. For example, the discovery of skeletons belonging to human ancenstors, such as Homo habilis, Homo erectus, and Homo floresiensis, provide indirect evidence for a theory of human evolution, since this theory can account for these observations.

We have already provided some indirect evidence for the ideology theory of oppression in chapter 4, when we showed that it can partly account for the persistent economic inequality in the US and the persistent gender inequality in Sweden. By showing that the economic inequality in the US and the associated beliefs of its citizens could partly be accounted for by the ideology theory of oppression, but not by the gunman theory, we provided some evidence for the former theory. The theory gained additional support when we showed that it could not only account for the persistence of gender inequality in Sweden, but also for men's stronger preference to succeed on the labour market, and the outcome of stylised bargaining games where the gender of the players is commonly known.

However, having shown that the theory is compatible with some set of observations should not be enough to convince us. After all, having shown that the theory is compatible with our observations of two persistent social orders is hardly conclusive evidence. In order to provide better evidence, we would have to expand our set of facts to include a greater number of oppressive social orders.

This brings us to what Kincaid calls direct evidence. This type of evidence consists of showing that the conditions of the functional explanation hold one by one. Kincaid suggests that this can be done with the same statistical methods economists use to show that economic theories hold.

In order to see this, let us recall that we formulated the ideology of oppression as follows:

> Oppressive societies maintain themselves without depending solely on coercion

1. in virtue of ideological beliefs and norms among a substantial part of the citizenry, and
2. these citizens are subject to ideology because this serves the function of upholding the oppressive status quo.

Establishing that condition 1 holds is, as we shall see, relatively straightforward. In order to make it easier to see how the functional claim in condition 2 can be established, let us reformulate it with the help of a Cohen-style consequence law as follows:

A. IF it is the case that ideology (at t_1) causes the persistence of the oppressive social order (at t_2), THEN ideology persists (at t_3).

B. Ideology (at t_1) causes the persistence of the oppressive social order (at t_2).

C. Therefore, ideology persists (at t_3).

If we manage to establish that the explanans, A and B, is true of an oppressive society *S*, then we will be able to functionally explain that ideology persists in *S*. Furthermore, since ideology causes the oppressive social order to persist, this will provide us with an explanation of why the oppressive social order, *S*, persists as well. In order to successfully establish that the explanans is true, we will have to establish that *S* has disposition B, and that consequence law A holds in *S*.

Before we move on to showing how this can be done let us first consider how we normally provide evidence for a theory in the social sciences. Consider the following explanation of increased gold prices:

D. Whenever there is financial instability, gold prices increase.

E. There is financial instability.

F. Therefore, gold prices increase.

In order to test whether law D holds, we have to look for cases of financial instability and investigate whether these have been associated with increased gold prices. Each case of financial instability associated with increased gold prices confirms the law, whereas each case of financial instability not followed by an increase in gold prices disconfirms it.

Similarly, we would have to establish that A and B hold in order to provide direct evidence for the ideology theory of oppression. To be more precise, we need to establish that 1) whenever a substantial part of the citizens are victims of ideology, the oppressive social order will persist (B), and 2) whenever this disposition exists in a society, ideology will persist (A).

In order to establish these claims, we need to identify the oppressive social orders where ideology and oppression persist. If we can do this, then we can attempt to establish whether ideology causes oppression by investigating the cases where a substantial part of the citizens are victims of ideology. Establishing this claim is analogous to establishing the law that connects financial instability with increased gold prices. Each case where both ideology and oppression is present will confirm the claim, and each case where ideology is present without oppression will disconfirm it.

At first sight, establishing the second claim might seem to be more complicated. However, in order to establish that consequence law A holds, we only have to follow the same steps as above.[1] We look for the cases where the major antecedent is satisfied, and then investigate whether the major consequent is satisfied in these cases. All cases where both the major antecedent and the major consequent are satisfied confirm the consequent law, and all cases where only the major antecedent is satisfied disconfirm it.

Since the major antecedent is satisfied by the social orders where ideology (among a substantial part of the citizens) causes the persistence of the oppressive social order, we have, by having established the first claim, already identified the relevant social orders. All cases where ideology persists after it has caused the social order to persist confirms the second claim. All cases where ideology does not endure after it has caused the oppressive social order to persist disconfirms the second claim.

The best way to conduct this type of investigation is with the help of statistical methods and a large data set. Consider, for example, how we can use statistics to investigate whether financial instability causes the gold price to increase.

Assume that we have a large set of time-sorted data about gold prices and financial stability. That is, we have information about gold prices and levels financial stability at times $t = (1, 2, \ldots, n)$. We can then use *regression analysis* to test whether financial instability causes increased gold prices. Or, to be more precise, it allows us to test whether financial instability is correlated with increases in gold prices. Although correlation does not imply causation, this is the best type of evidence we can get for a causal connection. We can, as we will see below, strengthen the evidence by attempting to rule out the interference of other potential causes. In the end, however, the best we can do is to show that there exists a strong correlation between two or more variables.

Regression analysis is a statistical method used to fit a function to actual observations. In our case, regression analysis allows us to describe gold price, A, as a function of financial stability, B, and regression coefficient, x, as fol-

[1] See also Cohen [2000, p 265].

lows:

$$A = Bx + \varepsilon.$$

A is called the dependent variable, in our case gold prices, and B is called an independent variable, in our case a measure of financial stability. x is a regression coefficient and ε is an error term. The closer the function is to the actual observations, the smaller will the error term be.

If an independent variable is causally linked to the dependent variable, then the corresponding regression coefficient will be significantly different from zero. The stronger the relationship between the dependent and independent variables, the more significant the regression coefficient will be. If there is no connection between the dependent and the independent variables, then the regression coefficient will not be significantly different from zero.

Furthermore, the error term picks up all information that is not captured by the independent variables. For example, assume that our data set includes observations of the gold market when it has been subjected to external shocks in form of a sudden increased supply of gold. Since we have not included external shocks as an independent variable, we have no regression coefficient that will account for these effects. The effect of the external shocks on the gold price will, therefore, be captured by the error term.

If the suggested claim that financial instability (low B) causes high gold prices (high A) is true, then our regression analysis will result in a negative x that is statistically significant from zero. If, on the other hand, financial instability is not connected to gold prices, then the regression coefficient will not be significantly different from zero. Furthermore, it is worth pointing out that although our lawlike statements are expressed as exceptionless generalisations, it is very seldom the case that we hit upon a function where the error term is zero and where all relevant information is captured by the regression coefficients. This can be seen as supporting the probabilistic interpretation of laws.

Although actual regression analyses are usually more complicated, this simple model suffices to illustrate how lawlike statements are tested in economics and the social sciences. Let us turn to the ideology theory of oppression and investigate whether we can use the same method to test its claims.

The first claim of the ideology theory was that ideology causes the persistence of the oppressive social order. Since this is analogous to the claim that financial instability causes high gold prices, we can formulate it as the following regression function:

$$O = Ix_1 + \varepsilon_1. \tag{7.1}$$

The dependent variable O is a measure of oppression in a social order, and the independent variable I is a measure of the prevalence of ideological beliefs

in a social order. If we accept that social orders with persistent substantial economic inequality are oppressive, then we can measure O with the help of, for example, the Gini index. The Gini index takes a number between 0 and 1 where more unequal societies score higher on the index. A perfectly egalitarian society will have a score of 0, and a perfectly unequal society, where one person owns everything, will have a score of 1. The independent variable, I, on the other hand, can be measured by, for example, the proportion of people who answer that they believe that their society has a higher degree of social mobility than it actually has. If the regression analysis results in a significant and positive x_1, then we will have some direct evidence for the claim that oppressive social orders maintain themselves in virtue of ideological beliefs among the citizens.

In order to test the second claim, we will need to formulate an appropriate regression function. Following a suggestion by Kincaid [1996, p. 116], we can represent the functional claim with the following regression function:

$$P = Ex_2 + \varepsilon_2. \tag{7.2}$$

The dependent variable, P, is a measure of the persistence of ideological beliefs in the population, and the independent variable, E, is a dummy that takes the value 1 if equation (7.1) held at an earlier point, and 0 otherwise. We test whether the fact that ideological beliefs caused inequality in a society at an earlier time, cause the persistence of the ideological beliefs at a later time.

Kincaid suggests some sophisticated methods that can be used to estimate the probability that a trait exists in a population after some initial state. A less sophisticated method of approximating the persistence of the ideological belief, P, would be to use the change of ideological answers during a time period. Using this measure for P and E, we can then run a regression. If the regression coefficient, x_2, is positive and significant, then we will have direct evidence for the second claim as well.

It is also worth pointing out that we can improve the evidence by including more independent variables. This can help us rule out the spurious correlations that Elster worried about. For example, recall that Tännsjö [2006] argued that both oppression and ideology were caused by the inability to solve coordination problems. In order to rule out this possibility, we could include the severity of coordination problems as an independent variable in both regression (7.1) and (7.2). Assuming that we can find a reliable way of measuring the severity of coordination problems, we can run a regression where we control for coordination problems. If our original regression coefficients, x_1 and x_2, remained significant after the inclusion of the new independent variable, then we could rule out that both oppression and ideology was caused by the inability to solve coordination problems.

It can now be asked that if there is such a simple method for testing the ideology theory of oppression, why have we not already tested it? The problem is that there is not yet enough relevant data. Although it is relatively easy to get access to data on inequality, unemployment, and other social indicators, it is much more difficult to get access to data on people's beliefs and values. We do not even have enough data on reported beliefs to run a reliable regression. Although we have some data, we do not have data collected over any longer period of time.

To see this, assume that we are interested in running regressions on (7.1) and (7.2). We would then need time-sorted data on both the Gini coefficient and reported beliefs about social mobility over time. Furthermore, for the regressions to be reliable we would need data from a long time period and from the same population. It is easy to gain access to time-sorted data on the Gini index. However, polls where people from the same population have been asked to report their beliefs about social mobility are not as common, and the few polls that have been conducted have not been repeated enough times to allow us to run a reliable regression.

However, it does *not* follow, from this somehow pessimistic observation, that it will never be possible to run the necessary regressions. The worldwide poll, *The World Values Survey*, did, for example, include the following question in their 1994 survey: "In your opinion, do most poor people in this country have a chance of escaping from poverty, or is there very little of chance escaping?"[1] Unfortunately, they have not asked it their later surveys. However, provided that they ask this (or some similar) question in future surveys, we should, in time, have enough data to test the ideology theory of oppression.

There may, of course, be other methods that demand less data that can be used to test the ideology theory of oppression. Finding such alternative methods will, however, be left as suggestions for future research. Nevertheless, we can conclude that it is possible to test the ideology theory of oppression with the same methods as we use to test other hypotheses in the social sciences.

7.3 Policy implications

Let us now turn to the question of what we ought to do if our empirical investigations show that the ideology theory of oppression is correct? That is, what shall we do if we discover an i) oppressive social order that persists because the oppressed are victims of ideological beliefs, and ii) the members of this society are victims of ideological beliefs because this serves the function of upholding the oppressive social order?

[1] See Robert Inglehart [2000].

As we pointed out in chapter 4, it is not necessary according to the ideology theory of oppression that everyone is a victim of ideology in order for the oppressive social order to maintain itself without depending solely on coercion. Let us assume that there is a small group of oppressed in an oppressive society who are not victims of ideological beliefs and norms, and ask what they should do if they want to exchange the oppressive social order for a less oppressive order.

First of all, if the ideology theory of oppression is correct, then the group of ideology-free revolutionaries cannot wait for the majority of the oppressed to spontaneously rise against the oppressive social order. After all, since everyone, except the small group, like things as they are, they will have to wait forever. In other words, the ideology-free group cannot be, as Lenin [1988, p. 140] accuses some of his competing revolutionaries, *subservient to spontaneity*. Instead they have to be, as Lenin puts it, the *vanguard* for the revolution and lead the oppressed in the struggle.

However, unlike in societies where the oppressed abstain from participating in revolutionary action because they fear punishment, the vanguard cannot increase the probability of a revolution by making punishment less likely. They cannot, for example, hope to mobilise the masses by providing means of coordination and communication, or by providing weapons and military training. After all, most of the oppressed do not perceive the social order as oppressive, and do not, therefore, believe that there is anything that needs to be changed.

For the same reason, the vanguard cannot hope to win an election where they make the (sincere) promise to, e.g., end poverty. After all, the citizens who are victims of ideology will not vote for a party that promises to fix something that does not, in their eyes, need to be fixed.

In order to gain the support of the oppressed, the vanguard will first have to get rid of the ideological beliefs and norms. One way of doing this is to organise awareness-raising campaigns in the hope that this will convince the ideology-ridden oppressed that the social order needs to be changed. This strategy has, for example, been advocated by women's rights activists in order to show women that what they thought were isolated, individual problems were in fact general problems facing all women.

It is interesting to note that Lenin [1988, p. 122] criticised the use of this strategy when it was focused on exposing the oppression of a single group at a time. According to Lenin, the sole focus on the condition for the workers in the factories has a chance to slightly improve their lot, but it cannot result in the abolishment of the oppressive system. To get the workers to participate in a revolution with the goal of changing the whole system, they have to be exposed to the oppression of all groups within the social order. Once they have realised that they are not the only ones who suffer under in the current social order,

they will cease to focus on their own narrow interests and come to support a revolution.

Lenin's critique of the use of narrowly focused awareness-raising campaigns gains some support from the ideology theory of oppression. In order for the campaign to have any long-term effects, it has to result in the abolishment of the oppressive social order. If it fails in achieving this, then the oppressive social order will keep on producing ideology. This brings us to the vanguard's next problem.

Assume that the vanguard engages in awareness-raising campaigns aiming to expose the oppressive character of the oppressive social order. With luck, such campaigns can rid some of the oppressed of their ideological beliefs and norms. However, if the revolutionaries hope to keep a substantial part of the oppressed free from ideology, they may become disappointed.

If the ideology theory of oppression is correct, then the oppressed are victims of ideology because this serves the function of upholding the oppressive social order. In other words, a substantial part of the oppressed will continue to be victims of ideology for as long as the oppressive social order persists. Although this does not mean that it is impossible to rid some individuals of their ideological beliefs through awareness-raising campaigns, it does mean that the influx of ideology will be greater than the outflow. For example, for each woman the women right's activists successfully free from ideology, at least one other woman will become the victim of ideology. The new ideology victim may be a girl born into an ideology-heavy subgroup of the population, or a former ideology-free woman who (for some reason or another) adopts the ideological norms.[1]

However, if we interpret the underlying consequence law of the ideology theory of oppression probabilistically, then the prospects of awareness-raising may not be as bad. Under this interpretation, the existence of the disposition does not ensure that ideology persist in the social order. Instead, its existence increases the probability that ideology will persist. So if it is possible for the vanguard to alter this probability with their campaigns, then they might have a chance to free the oppressed from ideology. It should, however, be pointed out that the vanguard needs to spend resources on demystifying their fellow oppressed, whereas ideological beliefs will be produced for as long as the social order is up and running. So even if the vanguard has a running chance of succeeding, it is likely that they will be fighting an uphill battle.

Assuming that it is not possible for awareness-raising campaigns to succeed, then what is to be done? The vanguard could always destroy the oppressive social order. After all, once the oppressive social order is gone there will

[1] See, e.g., Susan Faludi [2006].

be no more ideology victims. So, instead of first gaining the support of their fellow oppressed and then exchange the oppressive social order, the vanguard can first exchange the social order (against the expressed will of their fellow oppressed) and then gain their support. If the vanguard has correctly identified their social order as oppressive and their fellow oppressed are victims of ideology, then they do not have to worry about their fellows' complaints. After all, the reason why the other oppressed complain while the vanguard destroys the current social order is not that this is against their interests, but rather because they are victims of ideology. Once the revolution has succeeded, the oppressed will gratefully look back at what the vanguard has accomplished and recognise that the result was for the better.

Although it might follow from the ideology theory of oppression that any individual who wishes to change the oppressive system must engage in revolutionary activities against the expressed will of her fellow oppressed, it does not follow that the ideology-free oppressed should do so. Given what we know of the consequences of revolutions lead by small vanguards, it is also possible to draw a pessimistic conclusion. After all, if there is a significant risk that the revolution will involve horrible atrocities or that it will result in a prolonged reign of terror, then the oppressive social order might be preferable to the process leading up to the less oppressive social order.

However, just as the regression analyses have been left for the statisticians, and the hunt for the underlying micro-mechanism has been left for the sociologists, so will we leave the cost-benefit analyses of potential revolutions to social scientists, political philosophers, and would-be revolutionaries.

References

J.T. Adams. *The Epic of America*. Boston: Little, Brown and Co., 1931.

G. Ahrne and C. Roman. *Hemmet, barnen och makten (The Home, the Children and the Power). SOU 1997:139*. Statens offentliga utredningar. Stockholm: Fritze, 1997.

A. Al-Mohamad. Saudi women's rights: Stuck at a red light. *Arab Insight*, 2:45–52, 2008.

J.M.K. Alexander. *The Structural Evolution of Morality*. Cambridge: Cambridge University Press, 2007.

L. Althusser. *Lenin and Philosophy, and Other Essays*. New York: Monthly Review Press, 2001.

R. Aumann. Nash equilibria are not self-enforcing. In *Collected Papers*, pages 615–620. Cambridge, MA: MIT Press, 2000.

R. Axelrod. *The Evolution of Cooperation*. New York: Basic Books, 1984.

A.V. Banerjee. A simple model of herd behavior. *The Quarterly Journal of Economics*, 107:797–817, 1992.

G. de Beaumont and A. de Tocqueville. *On the Penitentiary System in the United States and Its Application in France*. Philadelphia: Carey, Lee & Blanchard, 1833.

G. Becker. *A Treatise on the Family*. Cambridge, MA: Harvard University Press, 1981.

C. Bicchieri. *The Grammar of Society*. Cambridge: Cambridge University Press, 2006.

C. Bicchieri. The fragility of fairness: An experimental investigation on the conditional status of pro-social norms. *Philosophical Issues*, 18:229–248, 2008.

C. Bicchieri and Y. Fukui. The great illusion: Ignorance, information cascades, and the persistance of unpopular norms. *Business Ethics Quarterly*, 9(1):127–155, 1999.

S. Bikhchandani, D. Hirshleifer, and I. Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *The Journal of Political Economy*, 100:992–1026, 1992.

S.J. Blackmore. *The Meme Machine*. Oxford: Oxford University Press, 2000.

S. Bromberger. Why-questions. In R. Colodny, editor, *Mind and Cosmos: Essays in Contemporary Science and Philosophy*. Pittsburgh: University of Pittsburgh Press, 1966.

K. Browne. *Biology at Work: Rethinking Sexual Equality*. New Brunswick, NJ: Rutgers University Press, 2002.

A. Buchanan. Revolutionary motivation and rationality. *Philosophy and Public Affairs*, 9:59–82, 1979.

M. Bunge. Mechanism and explanation. *Philosophy of the Social Sciences*, 27(4):410, 1997.

C. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press, 2003.

C. Camerer and R.H. Thaler. Anomalies: Ultimatums, dictators and manners. *The Journal of Economic Perspectives*, 9(2):209–219, 1995.

Central Intelligence Agency. North korea. In *The World Factbook 2009*. Washington, DC: Central Intelligence Agency, last accessed 27 April, 2012. URL `https://www.cia.gov/library/publications/the-world-factbook/`.

R. Chang. The possibility of parity. *Ethics*, 112:659–688, 2002.

G.A. Cohen. *Karl Marx's Theory of History: A Defense*. Oxford: Clarendon Press, 1978.

G.A. Cohen. Reply to Elster on "Marxism, functionalism, and game theory". *Theory and Society*, 11: 483–495, 1982.

G.A. Cohen. *Karl Marx's Theory of History: A Defense*. Oxford: Oxford University Press, 2000.

G.A. Cohen. Deeper into bullshit. In S. Buss and L. Overton, editors, *Contours of Agency: Essays on Themes from Harry Frankfurt*, pages 321–339. Cambridge, MA: MIT Press, 2002.

G.A. Cohen. *Rescuing Justice and Equality*. Cambridge, MA: Harvard University Press, 2008.

R. Cooper, D.V. DeJong, R. Forsythe, and T.W. Ross. Communication in coordination games. *The Quarterly Journal of Economics*, 107(2):739–771, 1992.

A.E. Cudd. How to explain oppression: Criteria of adequacy for normative explanatory theories. *Philosophy of the Social Sciences*, 35(1):20–49, 2005.

A.E. Cudd. *Analyzing Oppression*. New York: Oxford University Press, 2006.

R. Dawkins. Viruses of the mind. In B. Dalhbom, editor, *Dennett and his Critics: Demystifying Mind*, pages 12–27. Oxford: Blackwell, 1993.

R. Dawkins. *The Selfish Gene*. Oxford: Oxford University Press, 2006.

E. Durkheim. *Suicide: A Study in Sociology*. London: Routledge, 1979.

T. Eagleton. *Ideology: An Introduction*. London: Verso, 2007.

C. Eckel and P. Grossman. Chivalry and solidarity in ultimatum games. *Economic Inquiry*, 39(2):171–188, 2001.

T. Ellingsen and R. Östling. When does communication improve coordination? *The American Economic Review*, 100(4):1695–1724, 2010.

J. Elster. Cohen on Marx's theory of history. *Political Studies*, 28(1):121–128, 1980.

J. Elster. Marxism, functionalism, and game theory. *Theory and Society*, 11:453–482, 1982.

J. Elster. *Explaining Technical Change: A Case Study in the Philosophy of Science*. Cambridge: Cambridge University Press, 1983a.

J. Elster. *Sour Grapes*. Cambridge: Cambridge University Press, 1983b.

J. Elster. *Making Sense of Marx*. Cambridge: Cambridge University Press, 1985.

J. Elster. Norms of revenge. *Ethics*, 100(4):862–885, 1990.

S. Faludi. *Backlash: The Undeclared War Against American Women*. New York: Three Rivers Press, 2006.

E. Fehr and S. Gächter. Cooperation and punishment in public goods experiments. *The American Economic Review*, 90(4):980–994, 2000.

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

J.A. Fodor. Special sciences (or: The disunity of science as a working hypothesis). In M. Martin and L.C. McIntyre, editors, *Readings in the Philosophy of the Social Sciences*, pages 687–699. Cambridge, MA: MIT Press, 1994.

M. Foucault. *Discipline and Punish: The Birth of the Prison*. London: Penguin Books, 1991.

M. Friedman. The methodology of positive economics. In M. Martin and L.C. McIntyre, editors, *Readings in the Philosophy of the Social Sciences*, pages 647–660. Cambridge, MA: MIT Press, 1994.

A. Gibbard. *Wise Choices, Apt Feelings: A Theory of Normative Judgement*. Oxford: Clarendon Press, 1990.

A. Gibbard and H.R. Varian. Economic models. *The Journal of Philosophy*, 75(11):664–677, 1978.

S. Glennan. Rethinking mechanistic explanation. *Philosophy of Science*, 69(3):342–353, 2002.

J. Goldstone. Is revolution individually rational? *Rationality and Society*, 6:139–166, 1994.

S.J. Gould. The origin and function of "bizarre" structures: Antler size and skull size in the "Irish elk," *Megaloceros Giganteus*. *Evolution*, 28(2):191–220, 1974.

R. Hardin. *One for All: The Logic of Group Conflict*. Princeton, NJ: Princeton University Press, 1995.

S. Haslanger. Oppressions: Racial and other. In M. Levine and T. Pataki, editors, *Racism in Mind*, pages 97–123. Ithaca, NY: Cornell University Press, 2004.

P. Hedström and R. Swedberg. Social mechanisms. *Acta Sociologica*, 39(3):281–308, 1996.

P. Hedström and P. Ylikoski. Causal mechanisms in the social sciences. *The Annual Review of Sociology*, 36:49–67, 2010.

C.G. Hempel. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press, 1965.

C.G. Hempel. The logic of functional analysis. In M. Martin and L.C. McIntyre, editors, *Readings in the Philosophy of the Social Sciences*, pages 349–375. Cambridge, MA: MIT Press, 1994.

E.S. Herman and N. Chomsky. *Manufacturing Consent: The Political Economy of the Mass Media*. New York: Pantheon Books, 1988.

T. Hobbes. *Leviathan*. Cambridge: Cambridge University Press, 1996.

H.J. Holm. Gender-based focal points. *Games and Economic Behavior*, 32(2):292–314, 2000.

D. Hume. *A Treatise of Human Nature*. London: Penguin Books, 1985.

R. Inglehart. *World Values Surveys and European Values Surveys, 1981–1984, 1990–1993, and 1995–1997*. Inter-University Consortium for Political and Social Research Ann Arbor, MI, 2000.

M. Jiborn. *Voluntary Coercion: Collective Action and the Social Contract*. PhD thesis, Department of Philosophy, Lund University, 1999.

M. Jiborn. The power of coordination. Talk at Annual Conference of the Philosophical Society of Southern Africa, Grahamstown, South Africa, 2006.

G.S. Kavka. *Hobbesian Moral and Political Theory*. Princeton, N.J.: Princeton University Press, 1986.

H. Kincaid. Reduction, explanation, and individualism. In M. Martin and L.C. McIntyre, editors, *Readings in the Philosophy of the Social Sciences*, pages 497–513. Cambridge, MA: MIT Press, 1994.

H. Kincaid. *Philosophical Foundations of the Social Sciences: Analyzing Controversies in Social Research*. Cambridge: Cambridge University Press, 1996.

P. Kitcher and W. Salmon. Van Fraassen on explanation. *The Journal of Philosophy*, 84(6):315–330, 1987.

V. Lenin. *What Is to Be Done?* London: Penguin Classics, 1988.

S.D. Levitt and S.J. Dubner. *Freakonomics: a Rogue Economist Explores the Hidden Side of Everything*. New York: William Morrow, 2005.

T. Lewens. Cultural evolution. In E.N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, last accessed 29 April, 2012. URL `http://plato.stanford.edu/archives/fall2008/entries/evolution-cultural/`.

G.C. Loury. *The Anatomy of Racial Inequality*. Cambridge, Mass.: Harvard University Press, 2002.

G. Lukács. *History and Class Consciousness: Studies in Marxist Dialectics*. Cambridge, MA: MIT Press, 1972.

P. Machamer, L. Darden, and C.F. Craver. Thinking about mechanisms. *Philosophy of Science*, 67(1):1–25, 2000.

K. Marx. *Preface to a Critique of Political Economy*. London: Electronic Book Co., 2001.

K. Marx and F. Engels. *Critique of the Gotha Programme*. London: Electronic Book Co., 2001.

R.K. Merton. *Social Theory and Social Structure*. New York: Free Press, 1968.

R. Nagel. Unraveling in guessing games: An experimental study. *The American Economic Review*, 85(5): 1313–1326, 1995.

R. Nozick. *Anarchy, State, and Utopia*. New York: Basic Books, 1974.

M. Olson. *The Logic of Collective Action*. Cambridge, Mass.: Harvard University Press, 1971.

N. Olsson-Yaouzis. Revolutionaries, despots, and rationality. *Rationality and Society*, 22(3):283–299, 2010.

N. Olsson-Yaouzis. An evolutionary dynamic of revolutions. *Public Choice*, 151(3–4):497–515, 2012.

Organisation for Economic Co-operation and Development. *Economic Policy Reforms: Going for Growth*. OECD publishing, 2010.

E. Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press, 2005.

D. Parfit. *Reasons and Persons*. Oxford: Clarendon Press, 1987.

P. Pettit. The virtual reality of homo economicus. *The Monist*, 78(3):308–329, 1995.

P. Pettit. Functional explanation and virtual selection. *The British Journal for the Philosophy of Science*, 47:291–302, 1996.

G.W. Pierson. *Tocqueville in America*. New York: Anchor Books, 1959.

A. Przeworski. Material interests, class compromise, and the transition to socialism. *Politics and Society*, 10:125–153, 1980.

H.S. Pyper. The selfish text: The bible and memetics. In C.J. Exum and S.D. Moore, editors, *Biblical Studies/Cultural Studies: The Third Sheffield Colloquium*, pages 70–89. Sheffield: Sheffield Academic Press, 1998.

J. Rawls. *A Theory of Justice*. Oxford: Oxford University Press, 1999.

W. Reich. *The Mass Psychology of Fascism*. New York: Farrar, Straus & Giroux, 1970.

M. Resnik. *Choices: An Introduction to Decision Theory*. Minneapolis: University of Minnesota Press, 1987.

J.E. Roemer. *Analytical Marxism*. Cambridge: Cambridge University Press, 1986.

M. Rosen. *On Voluntary Servitude*. Cambridge, Mass.: Harvard University Press, 1996.

A. Rosenberg. *Philosophy of Social Science*. Boulder, CO: Westview Press, 2008.

B. Rothstein. The reproduction of gender inequality in Swe. *Gender, Work and Organization*, 2010.

W.C. Salmon. *Four Decades of Scientific Explanation*. Pittsburgh: University of Pittsburgh Press, 2006.

A.M. Savada. *North Korea: A Country Study*. Federal Research Division Library of Congress, 1993.

T.C. Schelling. *Micromotives and Macrobehavior*. New York: Norton, 1978.

T.C. Schelling. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press, 1980.

K.H. Schlag. Why imitate, and if so, how? A boundedly rational approach to multi-armed bandits. *The Journal of Economic Theory*, 78(1):130–156, 1998.

M. Silver. Political revolution and repression. *Public Choice*, 17:64–71, 1974.

T. Skocpol. *States and Social Revolution*. Cambridge: Cambridge University Press, 1979.

B. Skyrms. The stag hunt. *Proceedings and Addresses of the American Philosophical Association*, 75(2): 33–41, 2001.

S. Solnick. Gender differences in the ultimatum game. *Economic Inquiry*, 39(2):189–200, April 2001.

D. Sperber. An objection to the memetic approach to culture. In R. Aunger, editor, *Darwinizing Culture: The Status of Memetics as a Science*, pages 163–173. Oxford: Oxford University Press, 2000.

Statistics Canada. The wealth of Canadians: An overview of the results of the survey financial security. Research Paper, 2005.

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

L.W. Sumner. *Welfare, Happiness, and Ethics*. Oxford: Oxford University Press, 1996.

T. Tännsjö. Rational injustice. *Philosophy of the Social Sciences*, 36(4):423–439, 2006.

T. Tännsjö. Social psychology and the paradox of revolution. *South African Journal of Philosophy*, 26(2): 228–238, 2007.

M. Taylor. *Rationality and Revolution*. Cambridge: Cambridge University Press, 1988.

P.D. Taylor and L.B. Jonker. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1):145–156, 1978.

L. Thompson and A. Walker. Gender in families: Women and men in marriage, work, and parenthood. *Journal of Marriage and Family*, 51(4):845–871, 1989.

G. Tullock. The paradox of revolution. *Public Choice*, 11:88–99, 1971.

A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481): 453–458, 1981.

B.C. van Fraassen. *The Scientific Image*. Oxford: Oxford University Press, 1980.

J. Weibull. *Evolutionary Game Theory*. Cambridge, MA: MIT Press, 1997.

J. Weibull. What have we learned from evolutionary game theory so far? Working Paper Series 487, Research Institute of Industrial Economics, http://econpapers.repec.org/RePEc:hhs:iuiwop:0487, 1998.

A. Wertheimer. *Exploitation*. Princeton, N.J.: Princeton University Press, 1996.

E. Wolff. Recent trends in household wealth in the United States: Rising debt and the middle-class squeeze - An update to 2007. Working paper No. 589 - Levy Economic Institute of Bard College, 2010.

I.M. Young. *Justice and the Politics of Difference*. Princeton, N.J.: Princeton University Press, 2005.