Thinking about agreement : the empirical plausibility of moral contract theory

Timmerman, Jan

THINKING ABOUT AGREEMENT

The Empirical Plausibility of Moral Contract Theory

Peter Timmerman

THINKING ABOUT AGREEMENT

Thinking About Agreement. The Empirical Plausibility of Moral Contract Theory

ISBN: 978-94-6203-383-2

The research described in this thesis was funded by the Netherlands Organisation for Scientific Research (NWO).

The cover includes a picture of an election night crowd in Wellington, New Zealand. William Hall Raine took the picture in 1931. The Alexander Turnbull Library in Wellington keeps it as part of the collection 'Raine, William Hall, 1892-1955 :Negatives of New Zealand towns and scenery, and Fiji', with the reference number 1/2-066547-F. See the complete picture at http://natlib.govt.nz/records/22334852.

This book was printed in the Netherlands by Wöhrmann Print Service

RIJKSUNIVERSITEIT GRONINGEN

Thinking About Agreement

The Empirical Plausibility of Moral Contract Theory

Proefschrift

ter verkrijging van het doctoraat in de
Wijsbegeerte
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
donderdag 4 juli 2013
om 16.15 uur

door

JAN PIETER TIMMERMAN

geboren op 16 november 1984
te Diever

Promotores:	Prof. dr. M.V.B.P.M. van Hees Prof. dr. P. Kleingeld
Beoordelingscommissie:	Prof. dr. G.A. den Hartogh Prof. dr. A.J.M. Peijnenburg Prof. dr. M.M.S.K. Sie

Contents

1 Introduction 1

I *The Practicability Assumption*

2 Contract Theory and Perspective-Taking 25
3 Perspective-Taking in Moral Judgment 50
4 Perspective-Taking Accuracy and the Contract Test 85
5 How to Use a Contract Test 120

II *The Translucency Assumption*

6 Contract Theory and Translucency 143
7 Translucency and the Irrationality of Straightforward Maximization 163
8 Why Not Be an Opportunist? 196
9 When Constrained Maximization is Rational 228

10 Conclusions 239

Appendix 245
Samenvatting 254
Bibliography 261

Detailed Contents

1 Introduction

1 Introduction 1
2 Contemporary contract theory 5
2.1 Rawls 6
2.2 Gauthier 10
2.3 Scanlon 14
2.4 Conclusions 18
3 Two assumptions of contract theory 18
4 Overview of the book 22

I *The Practicability Assumption*

2 Contract Theory and Perspective-Taking

1 Introduction 25
2 The Practicability Assumption 26
3 Specifying the Practicability Assumption 29
3.1 Which agents should be able to apply the contract test? 29
3.2 When should agents be able to apply it? 32
3.3 After how much preparation should agents be able to apply it? 36
3.4 The Practicability Assumption specified 37
4 The contract test and perspective-taking 38
5 How to evaluate the Practicability Assumption 45
6 Conclusions 48

3 Perspective-Taking in Moral Judgment

1 Introduction 50
2 Two challenges 52
2.1 First challenge: moral judgments are typically caused by intuitions rather than reasoning 53
2.2 Second challenge: perspective-taking has no significant role in moral judgment 57
2.3 Conclusions 69
3 Support for a significant role for perspective-taking 69

3.1 Moral judgment and perspective-taking competence ... 70
3.2 Moral judgment and the exercise of perspective-taking ... 76
4 Conclusions ... 83

4 Perspective-Taking Accuracy and the Contract Test

1 Introduction ... 85
2 Identifying alternative standpoints ... 88
2.1 We often do not take other viewpoints sufficiently into account ... 89
2.2 Explaining egocentricity ... 91
2.3 Identifying all relevant standpoints ... 95
2.4 Conclusions ... 98
3 Understanding alternative standpoints ... 99
3.1 Egocentric understanding ... 99
3.2 Explaining egocentric interpretations ... 103
3.3 Conclusions ... 108
4 Contract test accuracy ... 109
4.1 Step 1: discerning the appropriate principle ... 110
4.2 Step 2: identifying alternative standpoints ... 111
4.3 Step 3: understanding the implications ... 112
5 Perspective-taking inaccuracy and the contract test's correct-usability ... 114
6 Conclusions ... 118

5 How to Use a Contract Test

1 Introduction ... 120
2 Gather information about alternative standpoints ... 123
3 Use the perspective-taking abilities of third parties ... 127
4 Internalise principles that would be the object of agreement ... 130
5 Is the Practicability Assumption empirically plausible? ... 137

II *The Translucency Assumption*

6 Contract Theory and Translucency

1 Introduction ... 143
2 The Compliance Problem ... 145
3 Constrained maximization and translucency ... 148
4 Constrained maximization versus straightforward maximization ... 153

5 Three challenges to the Translucency Assumption 158
5.1 First challenge: Translucency is psychologically implausible 158
5.2 Second challenge: People are not sufficiently translucent for constrained maximization to be more advantageous than straightforward maximization 159
5.3 Third challenge: Reserved maximization is more advantageous than constrained maximization 161
6 Conclusions 161

7 Translucency and the Irrationality of Straightforward Maximization

1 Introduction 163
2 Translucency in isolated interaction 165
2.1 Signs of trustworthiness 165
2.2 Translucency and cooperation 167
2.3 Judging trustworthiness on the basis of nonverbal information .. 170
2.4 Lie detection 173
2.5 Evaluating the Translucency Assumption for isolated interaction 175
3 Translucency in informed interaction 180
3.1 Translucency and previous observations 180
3.2 Translucency and third party judgments 183
3.3 Evaluating the Translucency Assumption for informed interaction 184
4 Interaction control and trust 187
4.1 Constrained maximization and interaction control 188
4.2 Straightforward maximization and interaction control 191
5 Conclusions 195

8 Why Not Be an Opportunist?

1 Introduction 196
2 The signs of reserved maximization 200
3 The risk of detection 204
3.1 Apparent golden opportunities 204
3.2 Limits of self-control 207
4 The fragility of trust 210
4.1 Reputation is easily cracked… 210
4.2 …and never well mended 213
5 Implications for reserved maximization 218

6 Reserved maximization and individual differences: the case of the successful psychopath 221
7 Conclusions 226

9 When Constrained Maximization is Rational

1 Introduction 228
2 What the findings do and do not show 228
3 How to get the most out of constrained maximization 232
3.1 Strive for optimality and against unfairness 233
3.2 Trust cautiously 234
3.3 Increase translucency 236

10 Conclusions

1 The empirical plausibility of moral contract theory 239
2 Some practical advice for contractarian agents 241

Appendix 245
Samenvatting 254
Bibliography 261

If one is concerned with morality for human beings, with the constraints that each can have reason to impose on herself, then we must face the limitations on human plasticity.

David Gauthier, 1988, p. 416

1

Introduction

1 Introduction

Moral theory is a normative discipline. Unlike psychology and other sciences that study how humans in fact behave, moral theory's primary question is how we *should* act. This normative question cannot be answered by empirical findings alone: *is* does not imply *ought*. But findings on human behaviour are relevant in another way. It is widely accepted that morality must be compatible with our nature. Morality must be within the reach of actual persons. *Is* does not imply *ought*, but *ought* does imply *can*. By putting forward obligations or ideals for which persons should strive, moral theories make assumptions about human abilities. Psychology and associated fields that study human nature can provide information about whether such assumptions are likely to be correct or not. Assumptions can either conform to the findings in these fields, in which case they are empirically plausible, or contradict them in some way, and thus be empirically implausible. A moral *theory* is empirically plausible only if its assumptions about human abilities are in accordance with empirical findings.[1]

Moral philosophers have always made substantial empirical assumptions, but these were seldom the object of systematic empirical scrutiny (Doris & Stich, 2011). This situation has changed dramatically in the last two decades. Researchers in the interdisciplinary field of moral psychology, including both psychologists and philosophers, have begun to examine the plausibility of ideas about our moral functioning in the light of empirical findings. This has led to debates regarding the empirical plausibility of several important

[1] As a theory may make additional assumptions about the world besides assumptions about human abilities, the latter being correct may not be sufficient for the theory to be empirically plausible.

approaches in moral theory, including virtue theory (Casebeer, 2003; Doris, 2002; Harman, 1999), utilitarianism (Greene & Baron, 2001; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008), and Kantianism (Berker, 2009; Greene, 2008; Mikhail, 2008). One approach that has, despite its recent rise to prominence, received little attention from moral psychologists is social contract theory. It is my aim to change this.

With the publication of *A Theory of Justice*, John Rawls reintroduced the idea of the social contract into contemporary moral thought. Rawls uses this idea to determine which social arrangements are justified. The crucial idea is that social arrangements are justified if and only if they would be the object of agreement among appropriately situated persons who are choosing their terms of interaction. Rawls's approach has had an enormous appeal, and the idea of the social contract plays an important role in the work of prominent theorists such as David Gauthier (1986), T.M Scanlon (1998), Stephen Darwall (2006) and Derek Parfit (2011). While these theorists interpret the idea of the social contract in different ways, they all hold that whether an action is right or wrong depends on whether the action, or more typically the *principles* that permit the action, would be the object of agreement.

Contract theory shares some of the psychological assumptions of other moral theories. Like Kantians and utilitarians, contract theorists assume that persons can form moral judgments through reasoning. Also, like Kantians and rule-utilitarians, contract theorists typically assume that persons can subsume individual actions or arrangements under general rules or principles. However, contract theorists also make certain assumptions not shared by other theorists. This is particularly so with regard to our abilities for thinking about others and their perspectives, on which I will therefore concentrate.

Contract theorists hold that to judge whether an action or institutional arrangement is morally justified, one must determine whether it is in conformity with principles that would be the object of agreement. They thus assume that persons are able to discern the content of this hypothetical agreement. They thereby assume, I will argue, that persons are able to determine the acceptability of principles from other perspectives than their own present point of view. This is one out of two assumptions on which my investigation regarding the empirical plausibility of contract theory will concentrate.

A second assumption is made, at least by some contract theorists, when explaining why we have reason to be moral; that is, why we have reason to perform actions and adhere to arrangements that are in conformity with

principles that would be the object of agreement. Some contract theorists argue that it is in each individual's interest to comply with such principles because otherwise one may be excluded from beneficial cooperative interactions. The assumption is that potential interaction partners can detect whether one can be trusted to comply or not, and as such will refrain from interacting cooperatively with persons who are not disposed to comply.

Both assumptions have been criticised. A persistent criticism of contract views is that, in contrast to what their defenders suggest, it is not clear what principles would in fact be the object of unanimous agreement (Braybrooke, 1987; Gauthier, 2003; Hare, 1973a). It is a well-known complaint against Rawls's contract theory, for example, that it is not evident that persons situated in the original position would come to agree on his two first principles. Whereas some critics only argue that they would agree on alternative principles (e.g. Harsanyi, 1975), other critics have argued that we are unable to discern what persons with wholly different perspectives from ourselves would come to agree on (e.g. Hare, 1973a).[2] They thereby question the plausibility of the first assumption.

The second assumption, which is mainly made by David Gauthier, has proved to be particularly controversial (Buchanan, 1990; Franssen, 1994; Nelson, 1988; Sayre-McCord, 1991). While we may expect that persons who violate moral principles will be detected now and then, it is far from evident that the likelihood of detection is so high that it is typically advantageous to comply. Many have found Gauthier's assumption that persons can be recognised as being untrustworthy even *before* they have committed a violation particularly unbelievable. In the words of one critic, "⌈p⌋eople cannot see inside each other's heads and it is idle to examine models in which they can" (Binmore, 1993, p. 138).

Neither proponents nor critics of these assumptions of contract theory have, however, turned to empirical evidence to make their case. Given that both assumptions concern our psychological abilities, empirical studies may help to resolve controversies such as these. I will in particular be concerned

[2] As Hare writes with respect to Rawls's contract theory: "The ⌈persons in the original position or⌋ POP game is in effect played by imagining ourselves in the original position and then choosing principles of justice. Rawls' POPs come to the decisions that they come to simply because they are replicas of Rawls himself with what altruism he has removed and a veil of ignorance clapped over his head. It is not surprising, therefore, that they reach conclusions which he can accept. If I myself play this game, I import into the original position *my* prejudices and inclinations, which in some respects are different from Rawls'. I have some inclination to insure against the worst calamities, in so far as this is possible. But I have no inclination to maximin, once the acceptable minimum is assured" (Hare, 1973, p. 249).

with studies in the field of *social cognition*. Psychologists use this term to refer to capacities and processes involved in perceiving, interacting with, and thinking about others. Social cognition is a central topic of research in psychology, and is also pursued by economists, neuroscientists, linguists, computer scientists, and logicians. There is much empirical work available which, I shall show, sheds light on the plausibility of the two assumptions, and thereby on the empirical plausibility of the social contract approach towards morality.

With this investigation, I hope to contribute to the philosophical debate in three ways. First and foremost, my aim is to contribute to the development of social contract theory. A theory cannot be plausible as a moral theory if it is not empirically plausible—if its assumptions about human abilities are not in accordance with empirical science. Secondly, the investigation will be of relevance for moral theory more generally. Most if not all moral theories make assumptions about our social cognitive capacities. As a diverse range of empirical studies will be considered, the study will also provide information about the plausibility of other assumptions than just the two on which it concentrates. Finally, the investigation will be relevant for the field of philosophical moral psychology. Whereas moral psychologists have debated about what empirical evidence on character, reasoning, and emotions show about our moral functioning, social cognition has received less attention. Moral cognition and social cognition are closely related, however: moral thinking usually concerns others. This study will bring to the fore empirical work relevant for understanding our moral functioning that has not yet received sufficient attention from moral psychologists.

This first chapter will be mostly concerned with describing contemporary contract theory. This will allow me to further explain the nature of the above two assumptions for contract theory and their role in contract theory. My discussion of contemporary contract theory in the following section focuses on the theories of Rawls, Gauthier, and Scanlon, who are the key figures in this tradition. Although Rawls's theory of justice is not itself an object of investigation, due to his influence on the moral contract theories of Gauthier and Scanlon it is helpful to take this theory as a starting point. In the third section I briefly return to the two assumptions; they will be further explicated in later chapters. The fourth section gives an overview of the book.

2 Contemporary contract theory

While the idea of a social contract can be found in the works of Thomas Hobbes, John Locke, and Jean-Jacques Rousseau, their use of it differs from that of contemporary contract theorists (D'Agostino, Gaus, & Thrasher, 2011; Kymlicka, 1993). Classical contract theorists used the idea of agreement mainly to explain political obligation. The central idea was that if citizens have (or could have) given their consent to the establishment of their government or to the laws of their government, then they have an obligation to obey government and abide by its laws. Contemporary contract theorists, on the other hand, use the idea of agreement to identify which political and social arrangements are justified. The crucial idea here is not that agreement generates obligation, but that it reveals "what we have reason to do in our social and political relations" (Freeman, 2007, p. 19).

After Rawls resurrected the idea of the social contract in the second half of the twentieth century, it was developed in different ways. Contract theorists are a diverse lot, including important thinkers such as Ken Binmore (2005), James Buchanan (1975), Stephen Darwall (2006), Samuel Freeman (2006), David Gauthier (1986), Jean Hampton (1993), John Harsanyi (1977), Gregory Kavka (1986), Jan Narveson (1988), Derek Parfit (2011), T.M. Scanlon (1998), and Nicholas Southwood (2010). It has become common to distinguish two strains in contemporary contract theory, one inspired by Hobbes and another more clearly influenced by Rousseau and Kant. This latter strain is often called *contractualism*, whereas the term *contractarianism*, although previously used to describe contract theory in general, is now typically reserved for the Hobbesian strain. While the central difference between these strains will play a role in the discussion below, I shall usually be concerned with contract theory in general. I will use the term 'contract theorist' to refer to thinkers of either strain who study and develop moral conceptions centred around the idea of agreement.

I shall in the following section explain contemporary contract theory further through a discussion of Rawls's theory. I then turn to the contract theories of Gauthier and Scanlon, as these will be the focus of my investigation. Gauthier is the most influential defender of Hobbesian contractarianism, while Scanlon is, after Rawls, the best known proponent of the sort that takes after Rousseau and Kant—he is also the one who coined the term 'contractualism'. A central difference between Rawls's theory on the one hand and Gauthier's and Scanlon's on the other is that Rawls concentrates on

political institutions, whereas Gauthier and Scanlon concentrate on interpersonal morality. Aspects of the theories of Gauthier and Scanlon most relevant for my purposes will be more extensively discussed in the later chapters.

2.1 Rawls

In *A Theory of Justice*, Rawls sets out to establish what it is for a society to be just. Society, Rawls says, can be thought of as an association of persons who in their relations to one another recognise certain rules of conduct as binding. These rules specify a system of cooperation with the purpose of advancing the good of those taking part in it. Society is thus "a cooperative venture for mutual advantage" (p. 4). But while all members of society are better off through social cooperation than if each relied solely on his own efforts, society typically also includes a conflict of interest. Persons are not indifferent as to how the benefits produced by their collaboration are distributed: each prefers a larger to a smaller share of the cooperative surplus. Put differently, they are not indifferent as to how the system of rules that constitute the basic institutions of society distributes the cooperative surplus—regarding what Rawls calls 'the basic structure of society'. A society therefore requires a set of principles which specify the basic structure of society. These principles constitute the society's conception of justice. But having a conception of justice, which is common to all societies, does not yet make a society just, Rawls emphasises. A society can only be just if it has the *appropriate* conception of justice, one that is *justifiable* to its citizens. Rawls's aim is to identify what principles constitute such a conception—to identify what he calls *the* principles of justice.

Rawls uses the idea of the social contract for this purpose. According to Rawls, the principles of justice are those principles that persons who are jointly determining their terms of interaction would agree to. In his words, "[t]hey are the principles that free and rational persons concerned to further their own interests would accept in an initial position of equality as defining the fundamental terms of their association" (p. 10). Rawls thus uses the idea of agreement to fix both the content and the rationale of the principles of justice: the principles that would be agreed to are the principles of justice and they are so because they would be agreed to. As he puts it, the idea is to settle "the question of justification … by working out a problem of deliberation" (Rawls, 1999, p. 16).

Rawls's use of the notion of agreement in this way is typical for contemporary contract theory in general. Contract theorists take the act of agreement to indicate what reasons persons have—they take it to be *reason-revealing* (D'Agostino et al., 2011). As D'Agostino, Gaus and Thrasher (2011) put it, "if individuals are rational, what they agree to reflects the reasons they have". If a principle or a social arrangement would be the object of unanimous agreement among rational persons, we know they have reason to accept it, and thus that it can be justified to them.

It is in part for this reason that Rawls does not rely on the idea of actual agreement but on that of *hypothetical* agreement. Principles of justice are principles that *would* be the object of agreement. Persons do not need to actually agree on a social arrangement to have reason to endorse it.

However, contemporary contract theories such as Rawls's are hypothetical in another way that may be thought to contradict the purpose of revealing our reasons. As we saw, Rawls does not hold that principles are justified if they would be agreed upon by us in our actual circumstances, but if they would be agreed upon by free, rational and equal persons who are choosing their terms of interaction. In general, contract theories take arrangements to be justified if they would be the object of agreement among idealised parties in an idealised agreement situation. It may be wondered how the choices of idealised parties can reveal what principles or arrangements we as actual persons have reason to endorse.

One of the main reasons to idealise is, however, precisely to ensure that we have reason to endorse the agreement.[3] Actual persons may be unaware or confused about considerations relevant for an agreement on the terms of interaction. They may lack relevant information, have false-beliefs, or be prone to bias. If these properties affect the content of their agreement, their agreement would not reveal those arrangements they have reason to endorse. For this reason contract theorists do not concentrate on an agreement between actual persons, but on an agreement between idealised representatives of actual persons.

Contract theorists typically assume the parties to the agreement to be both rational and to have all relevant information. Another idealisation that is usually made is that of mutual disinterestedness. The parties are concerned to

[3] Another important reason for idealisation is to have a *determinate* choice situation. Working out what rational persons would agree on is much less complex than working out what persons with various false-beliefs and biases would agree on. Indeed, Rawls justifies the veil of ignorance, to which I turn below, in part because it would not be possible for us take all particular facts into account when determining what the object of agreement would be.

further their own interest, and take no interest in the interests of others. This is in the first place to ensure that arrangements do not depend on feelings that may not be universal or stable. It is also to ensure that the agreement is not biased towards certain individuals. Some individuals may be more liked or more disliked than others. Furthermore, some individuals may assign more weight to the interests of certain others than vice versa. In order to avoid that individuals either profit or suffer unduly from such sentiments, contract theorists usually assume the parties to the agreement to have only self-regarding interests.

Rawls is well known for another idealisation that he makes with regard to the agreement situation. Rawls assumes that "no one knows his place in society, his class position or social status, nor does any know his fortune in the distribution of natural assets and abilities, his intelligence, strength, and the like" (1999, p. 11). The idea is that such knowledge may be expected to affect the sort of arrangements persons would prefer. Persons may use knowledge about their own situation to reach a more favourable agreement for themselves. In particular, more powerful parties may rely on the knowledge of their 'threat advantage' to extract better terms from those in worse positions (Freeman, 2012). To avoid this happening, Rawls denies persons in the agreement situation all knowledge of particular facts—they are placed behind a veil of ignorance. The veil of ignorance ensures that the parties are impartial. In effect, the agreement situation is fair between all the parties to the agreement—the parties are in "an initial position of equality" (p. 11). This fairness of the agreement situation, Rawls assumes, transfers to the principles chosen in it.

The idea of the veil of ignorance has been received critically by other contract theorists. In particular, many have contested that the veil of ignorance is required to generate impartial principles (Parfit, 2011). As we will see in the next two sections, neither Gauthier nor Scanlon follows Rawls in proposing the veil of ignorance as a condition for the agreement situation.

That other contract theorists diverge from Rawls with regard to the veil of ignorance makes clear that contract theorists may hold different views about the correct interpretation of the agreement situation. Indeed, as Rawls points out, the structure of contract theories can be divided in two parts: an interpretation of the agreement situation, and a set of arrangements that, it is claimed, would be agreed to by persons in that situation. The core aspects of Rawls's interpretation of the agreement situation, which he calls the *original*

position, have been described above. I will finish this section by briefly describing the relation between the first and the second part.

Rawls claims that persons in the original position would agree on two principles of justice. The first principle guarantees equal liberties for all. The second principle states that economic and social inequalities are only justified when they are attached to offices open to all and to the advantage of the worst off. While the exact arguments for these principles go beyond the purpose of the present text, it is worth noting that Rawls takes the principles to be selected through a process of individual rational choice: each person in the original position chooses this set of principles because it is the best way to secure her interests as a citizen in society.[4]

We may expect that when two contract theorists differ with regard to the first part of their theories, the interpretation of the agreement situation, they will also differ in the second part, the content of the agreement. In addition, we may expect that when two contract theorists have a similar interpretation of the agreement situation, they will arrive at similar conclusions with respect to the content of agreement as well. In line with this, Rawls has suggested that the above principles are 'deduced' from his original position. With respect to Rawls's theory, however, some authors have questioned whether the relation between the first and the second part is this straightforward (e.g. Hare, 1973a; Harsanyi, 1975). While concurring with Rawls's description of the original position, John Harsanyi has argued that parties would agree on the principle of average utility rather than the two principles of justice. But there may be a way to resolve the link between the first and the second part in this case. Although Harsanyi and Rawls accept a similar description of the original position, they hold different views with respect to the method of reasoning employed by the parties.[5] If we add this element to the first part of

[4] Some authors have taken this fact to imply that Rawls does not in fact have a contract theory (e.g. Hampton, 1980). Due to the veil of ignorance, there is no basis for bargaining in the agreement situation, and thus no real agreement, the argument goes. Against this, Freeman (2012) has pointed out that there is in fact an exchange between persons for something received: each agrees to commit to the principles of justice only on condition others do so too.

[5] The choice faced by the parties is a choice under condition of uncertainty. Rawls argues that rational persons in the original position employ a maximin rule of choice: they choose those principles that maximize the situation of the persons that are worst off. Rawls's reasoning is that, given the unique importance of the choice in the original position and given that they have no idea how likely they are to end up in the worst position of society, they want to make ensure this position is as good as possible. Harsanyi, on the other hand, argues that such persons would employ a principle of insufficient reason. Uncertainty does not lead them to choose as if they would be placed in the worst position, but leads them to assign equal probabilities to all positions in society.

their respective theories, they no longer derive a different agreement from the same agreement situation.

2.2 Gauthier

After Rawls's *Theory of Justice*, David Gauthier's *Morals by Agreement* (1986) is undoubtedly the next milestone in the development of contemporary contract theory. While clearly influenced by Rawls, Gauthier's use of the idea of the social contract diverges from Rawls in striking ways. The first difference concerns the scope of his theory. Unlike Rawls, Gauthier sets out to defend not only a political contract theory but also an ethical theory (Vallentyne, 1991a). Whereas Rawls aims to give us a conception of social justice, Gauthier aims to generate a complete conception of morality on the basis of the idea of agreement.

There is a second way in which Gauthier's theory is more ambitious than Rawls's. Rawls, as I have just mentioned, takes the principles of justice to be the object of rational choice from the original position. That does not mean that the principles are based in rationality alone, however. Rawls characterises the original position in such a way that, in his words, "the principles that would be chosen, whatever they turn out to be, are acceptable from a moral point of view" (p. 104). This is most clearly so with respect to the veil of ignorance: its function is to ensure that principles of justice are impartial with respect to social and natural inequalities, among other things. Rawls's original position is thus not free from moral presuppositions, and the principles of justice are consequently not derived from assumptions about rationality alone.

Gauthier rejects Rawls's approach. On his view, moral theory should resolve what he calls the "foundational crisis" of morality:

> From the standpoint of the agent, moral considerations present themselves as constraining his choices and action, in ways independent of his desires, aims, and interests. [...] And so we ask, what reason can a person have for recognizing and accepting a constraint that is independent of his desires and interests? He may agree that such a constraint would be *morally* justified; he would have reasons for accepting it *if* he had a reason for accepting morality. But what justifies paying attention to morality, rather than dismissing it as an appendage of outworn beliefs? (Gauthier, 1991a, p. 16)

Gauthier believes it is insufficient for a moral theory to only show what morality requires, as this would not show persons to have reason to be concerned with morality in the first place. He believes a moral theory must convincingly address what is usually called the 'Why be Moral?' question. To

answer this question, Gauthier aims to show morality is an effective way to further one's non-moral aims and interests. Or as he puts it, "to generate morality as a set of rational principles for choice" (Gauthier, 1986, p. 5).

This difference in aim has important consequences for Gauthier's interpretation of the agreement situation. In order to ensure that the morality agreed upon by the parties has a rational foundation, there is no place for a veil of ignorance that prevents the parties from being partial to their own interests. Persons in the agreement situation *must* have knowledge of their identities or they would be unable to see which arrangements would be rational for them to accept.

Having this knowledge obviously affects what arrangements the parties come to agree on. Whereas parties in the original position occupy, due to their uncertainty, a literally identical position with regard to the basic structure of society, parties in Gauthier's agreement situation are very much aware that certain arrangements are better for them than others. They thus favour different agreements. Gauthier claims that in order to settle on an agreement that is rationally acceptable for each, the parties engage in a bargaining process with one another in which each tries to get the best bargain available.

Gauthier envisions this bargaining process as follows. It begins from what he calls the initial bargaining position. Each of the bargainers has a starting point, which is supposed to represent "what she brings to the bargaining table" (1986, p. 130). Gauthier thinks of the initial bargaining position as involving a certain level of well-being or utility for each of the bargainers. They join the bargaining table in order to increase this initial bargaining position with the fruits of cooperation. What they have to figure out is, again, how to distribute the gains of cooperation—the cooperative surplus. Gauthier takes this procedure to involve two steps. First, each party advances a claim: the distribution that he would like others to agree to. As the claims made will typically be incompatible, concessions will have to be made until a set of mutually compatible claims is reached. This is the second step.

How do the bargainers reach an agreement? Gauthier argues that as each bargainer wants as much as possible of the cooperative surplus, in the first step of the bargaining process each claims as large a portion as possible—each person puts forward that agreement under which she does best. The size of this claim depends on that person's natural endowment. As Gauthier writes in an earlier paper, "[i]f we compare the well-being which accrues to the naturally intelligent, strong, and enterprising, under that arrangement maximally beneficial to such persons, with the well-being which accrues to the

naturally dull, weak, and lazy, under that arrangement maximally beneficial to them, we shall find the former to be greater" (1990, p. 163). Indeed, as he goes on to point out, the 'naturally gifted' may do better even under those arrangements maximally beneficial to the 'naturally deprived', as obtaining these may require larger rewards for the gifted.

In the second step the bargainers make concessions on their claims. Given their rationality and mutual disinterestedness, rational bargainers seek to minimise their concessions. Each will find it unacceptable to make a relatively larger concession than others, Gauthier claims, unless doing so is necessary for reaching agreement. According to Gauthier, this directs the bargainers to that agreement in which the greatest concession that any bargainer has to make with respect to his ideal agreement is the smallest. Gauthier calls this the Principle of Minimax Relative Concession.[6] This MRC principle serves as a rational starting point for further agreement on the terms of cooperation (Hampton, 1991).

Gauthier points out that the norms and practices that these contractors would agree to may not satisfy all our moral intuitions. In stark contrast with Rawls's original position, bargaining power plays a crucial role in Gauthier's agreement situation. A naturally talented person will put forward a larger claim than a naturally deprived person and will thus, after both have conceded a proportion of their claim, end up with a larger share of the cooperative surplus. A society based on the MRC principle may thus include much larger inequalities in wealth, responsibility, and power than a society based on Rawls's principles of justice. Moreover, agents without bargaining power may not obtain any rights under this account. As Gauthier puts it, "[a]nimals, the unborn, the congenitally handicapped and defective, fall beyond the pale of a morality tied to mutuality" (1986, p. 268).

There is, however, also an important part of our common sense morality that does fit the MRC principle, Gauthier claims:

> Many of our actual moral principles and practices are in effect applications of the requirements of minimax relative concession to particular contexts. We may suppose that promise-keeping, truth-telling, fairdealing, are to be defended by showing that adherence to them permits persons to co-operate in ways that may be expected to equalize, at least roughly, the relative benefits afforded by interaction. These are among the core practices of the morality that that we may commend to each individual by showing that it commands his rational agreement. (Gauthier, 1986, p. 156)

[6] I have left several steps out to keep the presentation simple. See Chapter 5 of Gauthier (1985) for the argument in detail and the exact content of this principle.

Gauthier holds that rational bargainers will come to agree on a crucial part of the practices and norms that make up our common sense morality. Given that the bargainers are idealised representatives of us, we also have reason to agree to these practices and norms.

However, by showing that there is a morality that it is rational to agree to, Gauthier has not yet shown that it is rational to *be* moral. For this to be the case, it must also be rational to *comply* with the agreement made. This is not evidently the case. While mutual cooperation is beneficial, it typically also involves costs for individuals. It is rational to agree to cooperate with another when these costs are exceeded by the fruits of cooperation. But if other parties to an agreement have already done their part, or if one's own behaviour cannot be observed by others, complying may not be rational. While it is certainly rational for me to agree to help another if he first helps me, it may not be rational to return the favour when the time comes. Gauthier argues that cooperative activities often have the structure of the Prisoner's Dilemma, in which defecting is almost universally taken to be the rational choice.

This Problem of Compliance, as it is sometimes called, is a thorny problem for contract theories that seek to derive morality from rationality. If moral requirements can contradict the demands of rationality, the contract theorist's claim that it is rational to be moral would clearly be false. Moreover, in that case we may not expect that rational bargainers in an agreement situation would come to agree on anything at all: why make an agreement with others if they cannot be expected to comply with its terms? Without a solution to this problem, then, the project of generating morality from rationality fails. As Gauthier puts it himself, in that case "we must conclude that a rational morality is a chimera" (1986, p. 158).

Gauthier believes he has a solution to the Problem of Compliance. He claims that it is rational to adopt moral norms as *constraints* on the pursuit of one's self-interest. When persons commit themselves to complying with the requirements of morality, they can gain each other's trust and cooperate to their mutual interest, Gauthier argues (Cudd, 2012). The idea behind this argument is that having the disposition to comply with moral norms affects how others respond to one. Gauthier assumes that a person's interaction partners can to a certain extent detect whether she is morally committed or not; he supposes that persons are what he calls *translucent* towards one another. Due to one's translucency, adopting moral constraint may be expected to have a positive effect on one's cooperative opportunities and is

therefore, Gauthier claims, rational. I discuss this argument extensively in Chapter 6.

Note that the Problem of Compliance does not arise for contract theorists such as Rawls who uses the idea of agreement only to explicate what morality or justice requires. With regard to situations in which it is not in one's individual interest to comply with the hypothetical agreement, such a theorist can consistently hold that even if compliance is not rationally required it is morally required. That is not to say, however, that a tension between self-interest and morality may not pose a problem for such a theorist. Many contract theorists, and certainly Rawls, are concerned with the *stability* of their moral conceptions. They suppose persons who live in a society governed by their conception will in general be motivated to comply with it; otherwise the conception would not be suited for its function. A large tension between morality and self-interest may reduce compliance to the point that a conception is insufficiently stable. I briefly return to this later (6§1).[7]

2.3 Scanlon

Gauthier, we have just seen, argues that a person has reason to treat others in accordance with principles that would be the object of agreement because it is in her own interest to do so. As he writes, "[t]he basic concern with agreement is to justify to *oneself* the constraints that adherence to moral principles requires" (2003, p. 168). As the title of his major work *What We Owe To Each Other* suggests, Scanlon holds a rather different view. Besides being motivated by their own interests, Scanlon takes persons to also be moved by a certain form of *respect* for others. This leads to a very different contract theory of interpersonal morality.

Scanlon's theory concentrates on the question of what it is for actions to be wrong. Put roughly, Scanlon holds that an action is wrong if and only if it is unjustifiable to others. More precisely, "an act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behaviour that no one could reasonably reject as a basis for informed, unforced general agreement" (p. 153). Note that this is different from saying that actions are wrong *because* they are unjustifiable; rather, it is to say that properties that make an action wrong are those properties that make the action unjustifiable. Or as Ashford and Mulgan (2012) put it, "[w]hat wrong acts have in common is that they cannot be justified to others".

[7] This means 'Chapter 6, section 1'. I will use this convention throughout the book.

The idea of an agreement situation is less explicit in Scanlon's contract theory than in that of Gauthier and Rawls. It is clearly there in the background, however. Moral principles are justified, Scanlon claims, if they would be the object of agreement among parties who seek principles that no one could reasonably reject as a basis for informed unforced agreement. Following Rawls and Gauthier, Scanlon takes the parties to the agreement situation to be informed, free, and mutually disinterested; and like Gauthier, he drops the veil of ignorance. Unlike both Rawls and Gauthier, however, Scanlon does not take the parties to be first and foremost concerned to advance their own interests. Instead, as the above description shows, Scanlon assumes that the parties are motivated to find an agreement that is acceptable to other parties who are similarly motivated.[8]

This 'motivational basis', as he sometimes calls it, is the distinctive aspect of Scanlon's theory. Scanlon attributes this motivation to the parties in the agreement situation because he takes it to be the most plausible interpretation of our concern with morality. As he writes, "those who are concerned with morality look for principles for application to their imperfect world which they could not reasonably reject, and which others in this world, who are not now moved by the desire for agreement, could not reasonably reject should they come to be so moved" (1982, p. 227). Acting only in ways that can be justified to others on the basis of terms they cannot reasonably reject is on Scanlon's view what morality is about; it is the 'subject matter of morality'. As he writes, "when we address our minds to a question of right and wrong, what we are trying to decide is, first and foremost, whether certain principles are ones that no one, if suitably motivated, could reasonably reject" (1998, p. 189).

Whereas Gauthier speaks of the rational acceptability of principles or arrangements, Scanlon speaks of their *reasonable* acceptability or, more frequently, reasonable rejectability. Whether a principle is reasonably rejectable or not depends on whether and what objections persons may pose against the principle. If the general acceptance of a given principle allowing some type of action imposes certain burdens on me, I have an objection to it. This does not yet mean I may reasonably reject the principle. Whether this is the case, depends on the burdens imposed by alternative principles that would replace the principle if it were rejected. Let's assume that in some particular

[8] In the original exposition of his contractualism, Scanlon writes that parties have "a desire [...] to find principles which none could reasonably reject insofar as they too have this desire" (1982, p. 243). Scanlon (1998) has however changed his view on the role of desires in both the explanation and the justification of action. He therefore now speaks more generally of 'being motivated' to find an agreement that no one can reasonably reject.

case the only alternative principle is one that *disallows* the type of action allowed by the principle that burdens me. If this alternative principle imposes an even greater burden on anyone else, I may not reasonably reject the principle under consideration, Scanlon claims. Indeed, if I am reasonable, I would redraw my objection once I see that the other has a stronger objection to the alternative (Ashford & Mulgan, 2012). On the other hand, if the alternative principle would not impose larger burdens on anyone else, I may reasonably reject the principle under consideration. This means that it would be wrong for others to perform actions governed by the principle.

It is worth noting that Scanlon holds that whether a principle imposes burdens on a person or not depends not on that person's preferences, but on "what people have reason to want" (p. 204). This includes reasons associated with their well-being, but goes further than just that. It includes reasons to avoid harm, to have our interests taken into account, to control our own bodies, to want outcomes to depend on our choices, to be treated fairly, to be able to rely on the assurances of others, and to give special attention to our own projects, friends, and family (from Frei, 2009; Scanlon, 1998; Scanlon, 2003). Note that Scanlon differs in this respect from Gauthier, who takes reasons to depend on individual preferences alone.

That Scanlon attributes a moral motivation to the parties in his agreement makes clear that he does not share Gauthier's concern with the 'Why be Moral?' question. Indeed, Scanlon thinks that, given that the great majority of people are in fact moved by moral concerns, moral theorists do not need to justify morality itself (cf. Freeman, 1991). What is needed, on his view, is "a fuller explanation of the reasons for action that moral conclusions supply" (Scanlon, 1998, p. 148). In particular, Scanlon thinks a moral theory must explain why moral reasons almost always take precedence over other reasons, and why we take it to be so important that people are morally motivated.

Scanlon does this in the first place by relating his account to the value of living with others on terms of mutual respect.[9] As he writes, the "ideal of acting in accord with principles that others (similarly motivated) could not reasonably reject is meant to characterize the relation with others the value and appeal of which underlies our reasons to do what morality requires" (p. 162). This relation, Scanlon says, is "worth seeking for its own sake". We have reason to want to stand in a relation of mutual respect with others. Given that

[9] In the discussion to which I refer, Scanlon speaks of a relation of mutual recognition rather than mutual respect. But in the secondary literature it has become more common to refer to it is as mutual respect.

respecting them requires treating them only in ways that can be justified to them, we have reason to do so.

Scanlon ties the ideal of justifiability also to the more familiar value of human life. The distinctive value of human life, he says, lies for an important part in the capacity to assess reasons and justification (1998, p. 105). Appreciating the value of human life must therefore involve recognising and respecting this rational capacity. To respond properly to this value, Scanlon proposes, we must treat rational others only in ways that would be allowed by principles that they could not reasonably reject.

Clearly, Scanlon has a rather different contract theory than Gauthier. Whereas Gauthier uses the idea of agreement to generate a mutually beneficial morality from non-moral premises, Scanlon's uses it to explicate certain moral ideals. As I mentioned before, writers on contract theory have come to take this difference to be so crucial that they place Gauthier and Scanlon in different contract theory traditions; Gauthier is the best known proponent of the more Hobbesian contractarianism, and Scanlon of the more Kantian contractualism (e.g. Darwall, 2003).

However, the distinction should not be taken too strictly. As D'Agostino, Gaus and Thrasher (2011) write, "there is often as much difference within these two approaches as between them". Moreover, there is much that they agree on. Both Gauthier and Scanlon use the idea of hypothetical agreement to explicate the content as well as the rationale of morality. They both conceive of morality as a system of rules that enables persons to have valuable relations with each other. And they both take moral persons to be motivated to justify themselves on terms that others have reason to accept. But whereas Gauthier characterises the relations that morality enables first and foremost by mutual advantage, Scanlon concentrates on mutual respect. And whereas Gauthier concentrates on the instrumental use of the disposition to justify oneself to others, Scanlon takes moral agents to have an intrinsic desire to justify themselves to others (Ashford & Mulgan, 2012).

The above discussion concerns the first part of Scanlon's contract theory, the interpretation of the agreement situation. What does the second part of his theory look like? Unlike Rawls and Gauthier, Scanlon does not identify one or two crucial major principles that subsequent arrangements must satisfy. To the contrary, he says that there is "an indefinite number" of valid moral principles (1998, p. 201). This reveals that Scanlon has a different focus than Gauthier and Rawls. Scanlon uses the idea of hypothetical agreement not in the first place to provide a justification of a particular conception of justice or

morality. Instead, he seeks to "characterize the method of reasoning through which we arrive at judgments of right and wrong" (p. 2). Indeed, as I understand Scanlon, he takes the idea of hypothetical agreement to play a central role in actual moral thought. Neither Rawls nor Gauthier seems to make this more descriptive claim.[10]

2.4 Conclusions

Despite several important differences, contract theorists are united by one fundamental idea: that moral principles are justified if and only if everyone has reason to agree to them. Contract theorists use the idea of hypothetical agreement as an instrument to identify these principles. The most fundamental difference among contract theorists concerns the question of what sort of reasons persons have for accepting or rejecting moral principles, which is reflected in their different interpretations of the agreement situation.

3 Two assumptions of contract theory

As I mentioned in the introduction, my investigation of the empirical plausibility of contract theory concentrates on two psychological assumptions made by contract theorists. Drawing on the above discussion, I will in this section describe these assumptions in somewhat more detail.

In order to avoid confusion, it may be helpful to briefly touch on a type of psychological assumption that my investigation will *not* concern. The above discussion mentioned several assumptions regarding the *parties* to the agreement situation. Parties are assumed to be rational, to be well informed, and to be mutually disinterested. They are supposed to be motivated to find an agreement that is best for themselves, or to find an agreement that none can reasonably reject. As I mentioned before, these assumptions are idealisations that are introduced to ensure that conclusions of the agreement situation have normative force for us. They are not meant to be generally true of *us*.[11]

---

[10] It is worth noting that Gauthier (1977) has in the past defended a similar descriptive claim. In particular, he has suggested that contract theory is part of the "deep structure of self-consciousness" of citizens in Western society (p. 326). Gauthier argues that citizens have come to conceive of their relations with each other in terms of contracts. The contractarian conception is, he writes, "gradually increasing its influence on our thoughts and leading us to abandon earlier ideas of human relationships as natural or supernatural rather than conventional" (p. 330).

[11] That is not to say that such idealisations cannot be empirically implausible. For an hypothetical agreement between idealised parties to be of relevance for us, the parties making them must not be too different from us (D'Agostino et al., 2011). Take for example the

Contract theorists are, however, committed to certain assumptions about human agents. I started this chapter by briefly mentioning the ought-implies-can principle. This principle implies that by proposing certain moral norms, contract theorists assume persons can comply with them. But the principle can be applied more generally. Contract theorists defend a particular type of moral conception that they claim we have reason to adopt.[12] They thereby suppose that persons *can* adopt this conception—that they can become *contractarian* moral agents.[13]

This empirical assumption includes a cognitive as well as a conative requirement. For agents to be able to rationally adopt a moral conception, they must first of all be able to *understand* the justification provided for the conception. They must be able to understand that the principles that the conception consists of would be the object of agreement. Contract theorists are committed to the assumption, or so I will claim in the next chapter, that agents are able to work out what would be the outcome of the agreement situation that they propose. Moreover, both Gauthier and Scanlon assume in addition that agents can use this idea to 'test' whether actions or norms are justified. Indeed, as we saw above, Scanlon seems to assume we already do so when we engage in moral reasoning. I will call the assumption that we can apply the idea of hypothetical agreement for moral evaluation the *Practicability Assumption.*

Adopting a moral conception involves being motivated to comply with its demands. By supposing that agents can rationally adopt a given moral conception, contract theorists thus suppose a conative requirement is satisfied: that agents can, or even will typically, *be moved* by its demands. Rawls, Scanlon, and Gauthier all make this assumption explicitly: they all argue their

---

assumption of mutual disinterest. As I mentioned before, contract theorists make this assumption for various reasons, one of which is to avoid the agreement being based on feelings not all of us have. Peter Vallentyne (1991b) has argued that this assumption may be problematic for a theory that seeks to show that a certain conception of morality is rational to accept, such as Gauthier's. Mutually disinterested persons may reach an agreement that would not be rationally acceptable for persons who *do* take an interest in the well-being of others, such as ourselves. For another example, consider Scanlon's assumption that parties are moved to justify themselves to others. While Scanlon does not claim that everyone has this motive nor that it is very strong in those who do have it, he does suppose it is part of what it is to be a moral agent (Ashford & Mulgan, 2012; Freeman, 1991). This reveals that through assumptions about the parties contract theorists may also commit themselves to assumptions about human abilities.

[12] My use of the term moral conception is based on Rawls's (e.g. Rawls, 1974). I take it to be in the first place a set of moral principles.

[13] I use the adjective 'contractarian' here and further on to refer to contract theory in general, including 'contractualist' theories.

conceptions of morality fit human motivation. Rawls claims that persons in the original position would choose his conception of justice in part because they take it to be more stable than alternative conceptions such as utilitarianism: the contractors recognise that citizens will, due to their sense of justice, generally be motivated to observe its requirements. Similarly, Scanlon assumes persons are motivated to justify themselves to others in a way that fits his moral conception.

Gauthier makes a stronger claim. Both Scanlon and Rawls suppose that persons accept certain moral ideals, and they rely on these to explain why persons would be motivated to comply with their moral conceptions. Gauthier, on the other hand, claims that agents can be motivated to comply with his moral conception solely on the basis of their non-moral interests. On Gauthier's view, being morally committed is an effective way to further one's non-moral interests. As I mentioned before, the idea behind this claim is that being morally committed may be expected to have a positive effect on how others respond to one. Persons are translucent, Gauthier assumes, and may therefore expect to do better for themselves by being moral. I will call this assumption the *Translucency Assumption.*

While Rawls and Scanlon do not make the Translucency Assumption, its plausibility is also relevant for their theories. As I shall explain later on (6§3), Gauthier's argument for the rationality of moral commitment is not essentially related to his particular moral conception. Roughly put, the idea is that it is advantageous to comply with norms that others expect one to follow, provided that the norms are mutually beneficial and not unfair. This idea can in principle also be combined with other moral conception than Gauthier's, such as Rawls's or Scanlon's.

Both the Practicability Assumption and the Translucency Assumption are empirical assumptions. Both assumptions state that persons have certain abilities. The Practicability Assumption states that persons have the abilities required to work out what others would agree to under certain conditions. The Translucency Assumption states that persons have the abilities required to recognise whether others are morally disposed or not. Both assumptions thus regard persons as having particular psychological abilities. More precisely, they state that we have social cognitive abilities: abilities to perceive and think about others. Empirical findings on our social cognitive abilities can thus reveal to what extent these assumptions are plausible.

My investigation regarding the empirical plausibility of contract theory will concentrate on the Practicability Assumption and the Translucency

Assumption. There are three main reasons for doing so. The first is that they are among the most crucial psychological assumptions in contemporary contract theory. If the Translucency Assumption would turn out to be false, Gauthier would have no response to the Compliance Problem. This would call into question his project of grounding morality in rationality. And if the Practicability Assumption would turn out to be false, the whole social contract approach may be in jeopardy. If persons do not have the capacities to work out what would be agreed upon in a hypothetical agreement situation, they would be unable to reliably derive moral principles on the basis of it. In that case, we would not be able to use the contract theorist's procedure for moral justification. This may also affect our reasons for accepting the arguments contract theorists give for certain moral principles—even if we would have an idea of what would be adopted by persons in a given agreement situation, we would know that our social cognitive abilities are such that we may not trust our judgment. Contract theorists may in that case still have an interesting idea of what needs to be the case for principles to be justified, but may not give us a moral conception that we can rationally adopt for living our lives.

The second reason for concentrating on these two is that it is far from clear that they are indeed plausible. As I mentioned in the introduction, philosophers have expressed their doubts about both of these assumptions. A quick glance at the empirical literature may strengthen such doubts. Psychological studies reveal that though we have capacities for thinking about others, or 'mindreading' capacities, we are far from excellent at it. As the social psychologist Nicholas Epley (2008) writes in a recent overview of the empirical findings, "people are fairly impressive mind readers in some instances and undeniably terrible in others" (p. 1456).

The third reason for concentrating on these two assumptions is that they both concern social cognitive capacities. As I mentioned in the first section, while there has been much research regarding social cognition in the past few decades, there has been little research regarding the implications of such findings for moral psychology and moral theory. Character, reasoning, and emotion, on the other hand, about which contract theorists may also make assumptions, have received more attention. An investigation into these two assumptions is therefore opportune.[14]

[14] By concentrating on these two assumptions I will not have done a complete investigation of contract theory's empirical plausibility. If they would turn out to be satisfied, it does not follow that contemporary contract theory is empirically plausible: there are other assumptions that require investigation. For one, I will not be able to investigate Scanlon's interesting empirical

4 Overview of the book

My investigation has two parts. The first part concerns the Practicability Assumption, the second part the Translucency Assumption. The parts can be read independently from one another.

The first chapter of Part I explains why contract theorists are committed to the Practicability Assumption. It will also put forward the claim that they thereby give the ability to consider other perspectives than one's own a crucial role in moral thinking. The two subsequent chapters consider what empirical findings on our capacity for perspective-taking show about the plausibility of the Practicability Assumption. Whereas Chapter 3 is mainly concerned with the role of perspective-taking in our moral thinking, Chapter 4 is concerned with how good we are at perspective-taking. In Chapter 5 I will discuss what we may conclude about the plausibility of the Practicability Assumption on the basis of the findings presented in the earlier chapters.

Part II starts with a chapter that explains in more detail why contract theorists may make the Translucency Assumption. The chapter introduces three challenges that have been posed against the Translucency Assumption, which are dealt with in the subsequent chapters. Chapter 7 considers first the challenge that people are not translucent at all, and then turns to the challenge that people are not sufficiently translucent for it to be rational to be moral. Chapter 8 discusses whether persons may not do better as opportunists who only act in accordance with morality most of the time, without actually committing themselves to morality. Finally, Chapter 9 considers what implications the empirical findings of the previous chapters have for Gauthier's argument for the rationality of being moral.

Chapter 10 presents my conclusions regarding the question of how empirically plausible moral contract theory is in the light of the findings on social cognition. I finish with some practical advice for those attracted to a contractarian conception of morality.

---

claim that persons are generally motivated to justify themselves to others. See also footnote 1 of this chapter.

I

The Practicability Assumption

Whosoever looketh into himself, and considereth what he doth, when he does think, opine, reason, hope, fear, &c, and upon what grounds; he shall thereby read and know, what are the thoughts, and Passions of all other men, upon the like occasions.

Hobbes, 1651/1991, p. 10

2

Contract Theory and Perspective-Taking

1 Introduction

Contract theorists defend a particular standard of moral justification: they hold that actions or practices are justified if and only if they conform to principles that would be the object of agreement. As I mentioned in the previous chapter, contract theorists assume agents can apply this standard as a 'test' to evaluate actions and practices. In the following chapters I shall investigate what empirical findings show about our ability to apply this contract test. In the present chapter I shall explain why this is of relevance and what approach I will take.

There are at least two reasons why it is of interest to investigate how good we are at applying the contract test. The first is associated with agents who are motivated to rely on it as a moral standard. An investigation into our ability to apply the contract test can help them to become better at doing so. The second reason is associated with the adequacy of the contract test as a procedure for justification. For the contract test to be an adequate procedure, it must be the case that people are able to apply it under the appropriate circumstances. I shall elaborate on this second reason in the following section.

This chapter will also be concerned with the question of how the ability to apply the contract test can be investigated. How able agents are at using the contract test depends, quite obviously, on what doing so requires of them. More precisely, it depends on what psychological capacities are required for applying it and to what extent agents have those capacities. Distinct procedures of moral justification may draw on different mental capacities or place different demands on them. Compare for example the principle of utility with Kant's categorical imperative. The principle of utility states that the right action is the action that maximises aggregate happiness. Applying this standard to assess the justification of actions requires a diverse range of

capacities, and most centrally a capacity to estimate the consequences of one's actions on everyone's happiness. Applying Kant's categorical imperative places different demands on our capacities. Kant states that actions are justified if and only if they are based on a maxim that one can, at the same time, will to be a universal law. While applying this criterion certainly involves a capacity to think of the consequences of actions, it does not require one to calculate aggregate happiness. It does, however, require some capacities not involved in applying the greatest happiness principle. For example, it requires the capacity to make explicit the maxim of one's potential action, as well as the capacity to think of this maxim as a universal law.

I will argue in the fourth section of this chapter that the contract test essentially involves a capacity to consider points of view different from one's own. To determine whether other persons would or could agree to a given principle one must consider the principle from their perspectives. If applying the contract test requires such a capacity, which is usually called *perspective-taking*, empirical findings on perspective-taking are relevant for examining how able we are at applying the contract test. While I shall leave the presentation of such findings for the following chapters, the fifth section of this chapter discusses in what way empirical findings may be relevant. This section also introduces two concerns with respect to our ability to use the contract test, which may either be confirmed or removed by a careful consideration of the evidence.[1]

2 The Practicability Assumption

The justificatory procedure that I refer to as the contract test has a central place in each contract theory. As I described in the previous chapter, contemporary contract theorists use the idea of hypothetical agreement for moral justification (D'Agostino et al., 2011). They use it to show that certain objects—be it actions, treatments, practices, policies, arrangements, or institutions—have certain normative properties, such as permissibility or justice. Contract theorists either hold that such objects have the relevant normative property if they would be agreed to under appropriate conditions,

[1] It is worth noting that the contract test is not the only procedure for justification to require a capacity for perspective-taking. Take for example the Impartial Spectator test that some utilitarians have defended. Determining which principles would be endorsed by an Impartial Spectator requires one to consider a perspective quite different from one's own. While I concentrate on the contract test, the following investigation may thus also provide relevant information regarding our ability to apply other reasoning procedures.

or, which is more common, if such objects conform to *principles* that would be agreed to. The moral contract theories of Gauthier and Scanlon are of the latter type: both hold that actions are justified if they conform to principles that would be the object of agreement under appropriate conditions.[2]

By proposing the contract test as an instrument for moral justification, contract theorists of course suppose it is adequate for that purpose. It should therefore satisfy the criteria that we may expect of such an instrument. A first criterion for such an instrument is *determinacy*: that when designated agents apply it under the appropriate conditions, it provides an answer. If the contract test is indeterminate, it would not be useful as a procedure for moral justification. A second criterion that we may expect a procedure of justification to satisfy is what I call *correct-usability*: that when designated agents apply the procedure under the appropriate conditions, they tend to do so correctly (i.e. without making mistakes). If the contract test would not satisfy this criterion it would not be a suitable justificatory device, as agents would have no reason to trust the conclusions they draw with it.

As I mentioned in the previous chapter, a recurring criticism of contract theorists is that their contract test is not in fact determinate.[3] This criticism sometimes concerns the *possibility* of hypothetical agreement.[4] It may also be *epistemological*, however: that even if there would be an agreement, we may not be able to determine its content.[5] In the following, I will suppose for the sake of argument that the contract test is determinate in the first sense. My interest lies with the epistemological question of whether persons are able to find out which principles would be the object of agreement. The criterion of determinacy concerns whether persons are able to draw conclusions about

---

[2] It may be worth noting that both Gauthier and Scanlon have used the term 'test' in this regard. Thus Gauthier (1991b) writes, "although we should not suppose that our actual moral practices and social institutions result from agreement, we may nevertheless hold that the appropriate justificatory test for the principles, practices, and institutions that govern and structure human interaction in ways that constrain the individuals involved is whether they would have been accepted by those individuals were they fully rational persons, each concerned to advance his own good (or the realization of his substantive aims), and collectively able to determine ex ante their terms and conditions of interaction by voluntary and unanimous agreement" (p. 324). And Scanlon writes, "the contractualist test of justifiability [explains] why failure to guard against harm in certain ways (but not others) is wrong" (Scanlon, 2003, p. 184).

[3] Both Braybrooke (1987) and Sugden (1993) have criticised Gauthier's theory in this respect, whereas Gauthier (2003) himself has argued that Scanlon's theory faces this problem.

[4] For example, Sugden (1993) claims that there may not be a unique solution to a bargaining interaction as envisioned by Gauthier.

[5] Braybrooke (1987) accuses Gauthier's contract theory of this type of indeterminacy when he argues that persons cannot determine which arrangements satisfy his contract test because they cannot know under which arrangements they would be best off.

which principles would be the object of agreement. I suspect that this criterion may well be satisfied: contract theorists at least themselves seem to be able to derive such conclusions. It is, however, another matter whether these conclusions are correct: whether the principles derived do indeed meet the contract test in that they would be the object of agreement under the conditions deemed appropriate. This is what the criterion of correct-usability is about.

I do not think that for the contract test to be an adequate instrument for moral justification it needs to be the case that it is always determinate and that a designated user always applies it correctly. No procedure of moral justification would satisfy these criteria to the fullest extent. I shall therefore say that the contract test can be used adequately by designated users if and only if these two criteria are satisfied to an appropriate extent. I do not have an exact measure for when this is so. Clearly, it should not be the case that the test tends to be indeterminate and that agents typically err when applying it. In addition, I assume it requires that the test is determinate and correctly usable with regard to certain important types of cases. I shall use such cases as a benchmark.

A crucial idea in the present investigation is that whether a procedure for justification such as the contract test can be used adequately depends to an important extent on the abilities of its intended users. They provide the 'hardware' on which the instrument must run. By proposing a procedure for moral justification, moral theorists commit themselves to psychological assumptions about its intended users. They must assume that its intended users are able to use the test adequately as a procedure of moral justification. This is what I have called the Practicability Assumption.[6]

The Practicability Assumption can be made more precise by distinguishing several variables. The first variable concerns *which agents* can use the procedure adequately to form moral judgments. Given that contract theorists propose their contract tests as procedures for real beings rather than ideal beings, they suppose that at least some of us are able to use them to form judgments about normative properties. This variable may, however, still range from including only a selected minority to including everyone.

A second variable, that has as yet remained implicit, concerns *the circumstances* under which agents can apply the procedure. This variable may

[6] Whereas I will be interested in the contract theorist's Practicability Assumption, I take it to be the case that any moral theorist who puts forward a procedure for moral justification is committed to such an assumption.

range from including only the relatively ideal circumstances under which philosophers engage in moral reflection to including also the more time-constrained circumstances of everyday moral practice. Clearly, under what circumstances a procedure for justification should be applicable depends in part on what its objects of evaluation are: whereas a procedure for evaluating the justness of societies does not have to be applicable in everyday situations, this may be different for a procedure intended to evaluate actions.

A third variable that we may distinguish concerns what we could call *the potentiality* of the agents' ability to apply the procedure. A proponent of a given procedure for moral justification may in principle assume that designated agents are *presently* able to apply it. However, they are more likely to only suppose that agents can *become* able to apply the procedure. Different suppositions may then be made about how far away they are from this point. They may be very close, only having to become aware of the distinct steps included in the test. But it is also possible that their present capacities are a long shot from what is required for applying the test adequately, and that much training is required to become an adequate user.

Given a particular procedure for moral justification, the Practicability Assumption is thus the assumption that designated agents are, after more or less preparation, able to apply this procedure adequately in the relevant circumstances. Contract theorists are committed to some variant of this assumption with regard to their procedure of justification, the contract test. The next section will specify the contract theorist's Practicability Assumption further.

3 Specifying the Practicability Assumption

The aim of this section is to specify the contract theorist's Practicability Assumption by considering how each of its three variables should be filled in by contract theorists. I will argue that moral contract theorists such as Gauthier and Scanlon assume that (1) actual agents can learn to apply the contract test (2) in the circumstances of everyday life (3) without too much difficulty.

3.1 Which agents should be able to apply the contract test?

Simply by publishing their contract tests as procedures for addressing questions of justification, contract theorists suppose they have an audience that can use it. Rawls, Gauthier, and Scanlon justify their respective moral

conceptions by means of a contract test. Rawls, for example, defends his particular conception of justice by arguing that his two principles of justice would be the object of agreement in the original position. In so far as contract theorists believe their readers can understand such an argument, they suppose them to be able to apply the contract test.

Of course, it does not follow from this that they take *everyone* to be able to apply the contract test. The readers of philosophy books form a set of individuals with relatively uncommon properties. In theory, a contract theorist could suppose that only this exclusive set of individuals can adequately use the contract test. Like Sidgwick, who suggested that none but the 'enlightened few' are able to fully understand utilitarianism, this contract theorist may deem it sufficient if only a few can use the contract criterion to address questions of justification. Indeed, if he thinks, as Gauthier (1977) once did, that public knowledge of the contractarian basis of society may lead to social problems, he may even intend that only an exclusive group of individuals can apply it adequately.

Most contract theorists assume not only that their readers can apply the contract test, however, but that *all* actual moral agents can come to do so.[7] This has to do with their view on the conditions a moral conception should satisfy. Contract theorists hold that a moral conception must be such that it can be rationally accepted by everyone, that it can play an effective social function, and that it can be fully public. I shall now briefly explain why each of these three conditions implies that actual moral agents must be able to learn to apply the contract test.

First, contract theorists hold that moral principles must be rationally acceptable for each of us. Gauthier, for example, writes on the first page of *Morals by Agreement* that he seeks to defend a morality that "is in each individual's reason" (1986, p. 1). Similarly, Scanlon holds that the basic characteristic of moral principles is that they are justifiable to everyone. As both authors hold that moral principles are justified if and only if they would be the object of agreement, they must assume that actual persons can in principle see whether or not a principle would be the object of agreement. That is to say, they must assume that persons can come to apply the contract test.

---

[7] Agents may have to satisfy certain standards of rationality or reasonability in order to be able to apply the contract test. The point for now is simply that their procedures of justification are not meant for ideal beings but for normal moral agents.

Second, contract theorists hold that it is a necessary condition of a moral conception that it can play an effective social or public function (Freeman, 2007). A moral conception should enable members of society to interact with one another on terms they can all rationally accept. This includes providing agents with moral principles as a shared basis for discussion, argument, and agreement. As Freeman (2007) writes, a moral conception should enable agents to "assess and criticize actions and institutions using shared criteria, and justify them to one another, when they are justifiable, on the basis of reasons all accept" (p. 6). It is worth noting that different contract theorists may provide a different rationale for the importance of public justification. Whereas Gauthier would argue it is of instrumental value for each individual as it enables mutually beneficial cooperation, Scanlon and Rawls would emphasise that mutual respect requires that we are in the position to justify ourselves to others on terms they can accept.

Part of the social function of a moral conception is that it must enable agents to assess and attribute accountability and responsibility. A moral conception should enable people to hold themselves and each other accountable. It is generally recognised that a person can only be held accountable for acting immorally if she is able to recognise the reasons for not doing so. Given that for a contract theory the justification of moral principles lies in their being the object of hypothetical agreement, a contract theory must hold that agents are able to recognise this and thus to apply the contract test.

The third and clearest reason why contract theorists require that actual moral agents can apply the contract test is that they believe a moral conception must be such that it can be public: that agents who are supposed to comply with the conception can know its justification (D'Agostino et al., 2011; Freeman, 2007). As a contractarian conception of morality is justified by being the object of hypothetical agreement, this requires that persons can apply the contract test. Contract theorists place importance on the publicity of a moral conception's justification for several reasons, two of which tie in with the above (for an extensive discussion on this condition see Freeman, 2007). First, as mentioned before, one of the essential ideas in the contract tradition is that a moral conception must be such that rational persons can freely accept the constraints that it imposes. Contract theorists therefore reject the idea of a moral conception that people would not accept once they knew its real justification.[8] Second, contract theorists hold that publicity is required for a

[8] Several contract theorists have argued that utilitarianism, at least as a conception of distributive justice, is a conception that does not satisfy this criterion (Rawls, 1999).

moral conception to be stable, in the sense that people will be moved to comply with its principles. People who cannot grasp and endorse the rationale of moral principles can hardly be expected to accept the constraints it proposes.[9]

I have argued that there are at least three reasons why contract theorists require that actual persons can understand which moral principles would be the object of agreement. The upshot is an answer regarding the first variable of the Practicability Assumption, regarding the question of which agents are supposed to apply the test: contract theorists assume that *actual persons* can come to apply the contract test adequately. (That is not to say, however, that they assume persons can *presently* apply the contract test to form moral judgments without any help or training; I return to this below.)

3.2 When should agents be able to apply it?

I now turn to the second variable of the Practicability Assumption, which concerns the circumstances in which persons must be able to apply the contract test. This turns out to depend on what purpose the test is to have in our moral lives.

One evident purpose of the contract test is for philosophical reflection on the justification of actions and practices. This is the way in which one uses Rawls's test when attempting to understand what a just society looks like, or when one uses Scanlon's test to determine how much persons ought to give to charity. Such questions can be addressed under circumstances in which there is ample time and opportunity to access certain relevant information and consider all the details. I can take another look at the relevant pages in these authors' books, or call for the advice of other parties more experienced than I am.

Justificatory problems may, however, also arise in everyday moral life. Persons continuously have to form judgments about what they have reason to do; to judge whether actions they may perform are morally permissible or not. They have to justify their own behaviour to others, or to assess whether the behaviour of others was justified. Contract theorists may assign their contract tests the purpose of solving such problems as well. I will say that contract theorists in that case propose the contract test as a *moral guide*.

---

[9] Gauthier, for example, takes publicity to be required because otherwise individuals cannot count upon their peers complying with moral arrangements, without which they have no reason to comply themselves.

If the contract test is proposed as a moral guide, agents must be able to apply it under conditions that are typically less ideal than those of philosophical reflection. In our interactions with others we often only have a short period to make up our mind and decide what to do. Moreover, we are typically ill-informed when choosing our actions, and under the influence of "stress and temptation" (Hare, 1973b, p. 153). The circumstances of philosophical reflection and those of day-to-day moral judgment differ in a way that may reflect how able people are to use the contract test under such circumstances. Whereas during philosophical reflection there are no evident constraints on the time, attention, and effort that can be put into using the contract test, there are such constraints in situations of everyday life. In one word, there is a difference in available cognitive resources.[10]

I take moral contract theorists such as Gauthier and Scanlon to be committed to the assumption that the contract test can be relied upon as a moral guide and thus be applied in everyday situations.[11] Moral contract theorists use the idea of hypothetical agreement to identify our obligations towards each other. They claim that our actions should be in conformity with principles that would be the object of hypothetical agreement. By ought-implies-can, they must then also hold that agents are able to act in conformity with such principles. For agents to be able to consistently do so, they must also be able to *assess* whether possible actions are in conformity with such principles or not. As such justificatory problems come up in situations of everyday life, agents must be able to apply the contract test in such situations.

Remarks by Scanlon and Gauthier suggest they indeed assume persons can use the contract test for moral judgment in everyday situations. Scanlon (1998) writes that "in order to decide whether it would be wrong to do X in circumstances C, we should consider possible principles governing how one may act in such situations, and ask whether any principle that permitted one to do X in those circumstances could, for that reason, reasonably be rejected (p. 195). This remark and many others like it suggest Scanlon's contract test can be used to make decisions in everyday life. Similarly, Gauthier (1986) writes that "the narrowly compliant [that is, moral] person [...] is prepared to be co-operative whenever cooperation can be mutually beneficial on terms equally

[10] While time is not a cognitive resource, it affects how many cognitive resources are available for using the contract test.

[11] More so than Gauthier, Scanlon is relatively explicit about this. When comparing his own theory with Rawls's, he writes: "there are important differences between the subject of Rawls's theory and the one being considered here. To begin with, Rawls's principles of justice are not intended to guide every choice and policy" (1998, p. 228).

rational and fair to all" (p. 179). Given that on Gauthier's view terms are rational and fair if and only if they approximate terms that would be the object of rational agreement, and assuming it will not always be evident in practice that possible terms of interaction are rational and fair, this presupposes that persons can apply his test in everyday situations.

It is important to emphasise that this does not imply that the contract test is a *decision procedure* that agents need to apply consciously *whenever* they make choices. As Mill already pointed out, behaviour may satisfy a moral standard without being based on it. A person may choose in conformity with principles that satisfy the contract test without choosing on the basis of it. However, a person is highly unlikely to *consistently* satisfy this standard if he is unable to assess whether his actions conform to principles that would be the object of hypothetical agreement. Persons do not need to be able to use the contract test as a decision procedure in order to live according to a contractarian moral conception, but it must be available to them as a guide.

It may be objected at this point that if the contract test is first and foremost a guide to examining moral principles and not a decision procedure, persons do not need to be able to apply it in everyday life. It would be sufficient if persons can use the test to reflect on moral principles in a cool hour, to borrow a phrase of Hare's, internalise principles that satisfy the test, and rely on these to evaluate actions or practices in everyday situations.

Although I endorse the idea that persons can learn and internalise conclusions of the contract test so as to economise on cognitive resources (3§2.1), it is unlikely that they can fully prepare themselves for moral practice from a cool hour. First, the moral principles that persons may adopt in such moments will be insufficiently detailed to cover all the situations in which they will find themselves. In order to be memorable, moral principles that would be adopted in such a cool hour must be verbally succinct and general. However, moral situations are often too complex to be assessable by such principles. As Scanlon writes:

> even the most familiar moral principles are not rules which can be easily applied without appeals to judgment. Their succinct verbal formulations turn out on closer examination to be mere labels for much more complex ideas. (Scanlon, 1998, p. 198)

Second, given the wide range of situations in which persons may find themselves, they cannot from a cool hour adopt all the principles that they are going to need in practice. As I mentioned before, Scanlon points out that there is an "indefinite number" of valid moral principles. I take this to imply that for

persons to be able to ensure that their actions are in conformity with principles that satisfy the contract test, they must at least *sometimes* consider the justification of such principles in everyday situations. They must thus be able to apply the contract test in everyday situations.

In order to find out to what extent agents can rely on the contract test as a moral guide, I will assess under what circumstances they can apply it. For expository purposes, I shall distinguish between three types of circumstances on the basis of available cognitive resources; two extremes, and one in the middle. The one extreme is circumstances of high cognitive resources. Our cognitive resources are high when we have plenty of time and attention to reflect. When a student uses the contract test in the course of writing a paper about the question of whether one ought to give to charity, she does so under circumstances of high cognitive resources. The student would not be under direct pressure of time when using it and can thus carefully consider all the relevant details. The other extreme is circumstances of low cognitive resources. If I were to use the contract test to judge whether I should give money to charity after just having been asked to do so by a collector on my doorstep, I would be using it under conditions of low cognitive resources. I have only a short time for making my decision, as the collector is waiting for my response. The third type of circumstance that I shall distinguish is that of medium cognitive resources. If a person in the course of his day uses the contract test to think about whether to give to a charity from which he received a donation letter, he would be using it under circumstances of medium cognitive resources. He has some time to think about it and can consider some details, but the resources that he spends on it are limited by all the other things he has to think about and do.

I take situations of high cognitive resources to be uncommon for most of us. Through the day we have to fulfil a large number of aims and goals, and have little time or cognitive resources to stand still and reflect on moral principles. We are usually surrounded by others, who are unlikely to leave us uninterrupted for long. We may have short periods for reflection when we drive to work, when we wait in the elevator, or when we answer an email message, but due to the constraints of everyday life I would say that such a situation only involves medium cognitive resources. Everyday life, I take it, includes for most of us few situations of high cognitive resources. Our available cognitive resources are usually either medium or low. For the rest, when I speak of everyday situations I will be referring to situations involving either medium or low cognitive resources.

3.3 After how much preparation should agents be able to apply it?

That brings us to the third variable of the Practicability Assumption, which I called in the previous section the potentiality of agents' ability to apply a procedure for moral justification. How much training does it require for persons to become able to apply the contract test adequately?

It is interesting to note first that Scanlon at times suggests persons are *already* able to apply the contract test.[12] More precisely, some of his phrases suggests he thinks persons already *actually use* it sometimes. Scanlon writes that his aim is to "characterize the method of reasoning through which we arrive at judgments of right and wrong" (p. 2). When he gives his answer to this question later on in the book, he writes that "when we address our minds to a question of right and wrong, what we are trying to decide is, first and foremost, whether certain principles are ones that no one, if suitably motivated, could reasonably reject" (p. 188). As I mentioned before (1§2.3), statements such as these suggest that Scanlon does not just propose a normative theory, but also, as he calls it himself, "an account of moral thinking" (p. 6). This interpretation is supported by the fact that Scanlon often appeals to commonalities between his contract test and our actual moral thinking. For example, Scanlon argues extensively that the ideal of being able to justify ourselves to others, which is the basis of the contractualist test, plays an important role in our moral experience and our practical reasoning (pp. 155-158).

Despite these suggestions, contract theorists do not require that persons can already apply the contract test. They require, and so does their Practicability Assumption, that agents can *learn* to use the contract test for moral justification, but it is not a problem if they require some practice before being able to do so adequately. Note, however, that it may be a problem if this learning process turns out to be very difficult or effortful or difficult. For one, persons may in that case not be sufficiently motivated to adopt the contract

[12] Although in *Morals by Agreement* Gauthier does not suggest that his justification of morality fits actual moral thinking, he does so in his earlier essay 'The Social Contract as Ideology'. Gauthier argues there that contract theory is part of the "deep structure of self-consciousness" (Gauthier, 1977, p. 326). He claims that citizens in Western society have come to conceive of their relations with each other as contractual. The contractarian conception is "gradually increasing its influence on our thoughts and leading us to abandon earlier ideas of human relationships as natural or supernatural rather than conventional" (p. 330). By doing so, it affects our practices and the arguments that we accept. Indeed, Gauthier argues that persons more and more think of their obligations towards one another in terms of contractual agreements.

test as a standard of conduct, such that the associated moral conception would in turn not gain sufficient grounding to fulfil its social role. I shall therefore take the Practicability Assumption to state that persons can learn to adequately apply the contract test without too much difficulty.

This is, however, not to say that the question of how able persons presently are at applying the contract is irrelevant for evaluating the Practicability Assumption. If a person would at present be completely unable to apply the contract test, we have no reason for thinking that he can, without too much difficulty, become able at applying it. More precisely, if a person would not have the sort of *general psychological capacities* required for applying the contract test, we have no reason for believing he can without too much difficulty adopt it as a moral guide. To the contrary, given that human cognitive plasticity is limited, the finding that persons are lacking in abilities that are crucial for applying the test may indicate that they *cannot* learn to apply it adequately. As I will explain further in §5 of this chapter, I take this to mean that the empirical plausibility of the Practicability Assumption can be assessed on the basis of studies regarding our present psychological abilities.

I shall in the following two chapters discuss to what extent empirical findings regarding our present abilities suggest that we can apply the contract test. The learning aspect of the Practicability Assumption will thus not play an explicit role in these chapters. For the sake of brevity, I will therefore typically not refer to this aspect when stating the Practicability Assumption. Nevertheless, it is important to keep in mind that the Practicability Assumption does not in the end concern our present abilities—these are relevant only in so far as they are indicative of our potential abilities. The learning aspect of the Practicability Assumption will be prominent in Chapter 5, as I will there consider to what extent persons can improve their ability for applying the contract test.

3.4 The Practicability Assumption specified

By using the contract test to justify their moral conceptions, contract theorists presuppose that persons are able to apply the contract test. Moreover, in so far as they propose it as a moral guide for dealing with justificatory problems that arise in everyday life, contract theorists must suppose persons can also apply it under less ideal conditions. I have argued that moral contract theorists such as Scanlon and Gauthier are committed to this assumption. More precisely, they are committed to the assumption that *actual persons can, without too much difficulty, learn to apply the contract test adequately under circumstances that include*

those typical of everyday life. As this is a cumbersome phrase, I shall often shorten it to 'that persons can apply the contract test' while meaning the same. In the following chapters I will consider to what extent this variant of the Practicability Assumption is plausible in the light of empirical findings.[13]

4 The contract test and perspective-taking

Contract theorists are committed to the assumption that persons can apply the contract test, I argued in the previous section. This implies that persons must have the capacities required for applying such a test. Critical among these is the capacity to consider alternative points of view, I shall argue in this section.

I shall in the following discussion focus on Scanlon's moral contract theory, as he is most explicit in assuming that this capacity is involved in applying the contract test.[14] He writes, for example, that we have "a direct reason to be concerned with other people's points of view: not because we might, for all we know, actually be them, or because we might occupy their position in some other possible world, but in order to find principles that they, as well as we, have reason to accept" (1998, p. 191). After having clarified the role of perspective-taking in Scanlon's contract test, I will briefly argue why this capacity is also crucial for other contract tests, including Gauthier's.

Scanlon holds that to judge whether an action is right or wrong is to judge whether principles that would allow it would be objects of agreement. More precisely, he holds that it is to judge whether principles that would allow it could or could not reasonably be rejected by people who were moved to find principles for the general regulation of behaviour that persons similarly motivated could not reasonably reject. As he writes: "in order to decide whether it would be wrong to do X in circumstances C, we should consider possible principles governing how one may act in such situations, and ask

[13] Note that if this variant of the Practicability Assumption turns out not to be satisfied, this would not imply contract theory fails as an approach in moral theory. It would, however, have implications for its ambitions. Say we were to find that the contract test can only be applied adequately by specialised individuals under relatively ideal circumstances. In that case, contract theory may still propose a method that such specialists may use to reflect on the justification of institutions and practices. But it would be less suitable as a theory for our obligations towards one another. Most persons would be unable to rationally adopt the moral principles that contract theorists deem justified. And even those who are sufficiently specialised to apply the contract test in a cool hour would, given the large variety of moral situations, sometimes not be able to find out whether concrete actions or practice are justified or not.

[14] That we must rely on perspective-taking to apply Scanlon's contract test is also pointed out by Darwall ('Introduction' in 2003).

whether any principle that permitted one to do X in those circumstances could, for that reason, reasonably be rejected" (1998, p. 195). This requires us to consider what Scanlon calls 'the burdens' that such possible principles impose on ourselves and others. With 'burdens' Scanlon refers to the implications of the principle being accepted, including consequences of the action X, that give those affected reason to object to the principle. Scanlon has recently described this as engaging in "a series of thought experiments, corresponding to the various ways that people might be affected by the principle in question" (2011, p. 132). Besides the burdens of general permission, we must also consider the burdens associated with agents *not* being permitted to do X in circumstances C—or as Scanlon puts it, "we need to consider the ways in which others would be burdened by a principle forbidding one to do X in these circumstances" (p. 195). Scanlon says that if we find that the burdens imposed by permitting one to do X in C are significantly greater than the burdens imposed on others by forbidding one to do X in C, we should conclude that any principle that permits X in C may reasonably be rejected by those affected by the greater burdens and thus that the action is wrong. Alternatively, if there were some principle for regulating behaviour in such situations that would permit one to do X and that would not burden others to such an extent that they can reasonably reject the principle, "doing X would not be wrong: it could be justified to others on grounds that they could not reasonably refuse to accept" (p. 195).

We already saw that Scanlon takes this procedure to involve perspective-taking. As he writes at another point, "we have reason to consider whether there are standpoints other than our own present standpoint from which the principles we are considering could reasonably be rejected" (p. 202). From the moment that we consider how others would be burdened by a principle, we consider the principle in question no longer solely from our own point of view, but from the points of view of others who are affected by it. We engage in thought experiments that are centred on positions regarding the principle other than our own, with the aim of finding out whether a person occupying that position has reason to agree to it.

Applying the contract test may sound like a complex process. Indeed, it may seem unlikely that persons in everyday situations are usually able to adequately go through this process. It is important to emphasise, however, that the above characterisation is not meant as a psychological description of the reasoning process that an agent must consciously, step-by-step, go through when applying the contract test. It is a rational reconstruction that

makes explicit what must, somehow, be taken into account when applying the test. We may not need to consciously apply the procedure as it is described above in order to find out whether principles or actions satisfy it or not. To be sure, it seems that often when we turn our mind to actions that are wrong in the sense that they are only allowed by principles that are unacceptable for certain others, we can recognise this directly, without having to engage in explicit reasoning. However, in order to recognise this, we must have a sense of the perspectives of these others: we must have a sense of the implications that the action or principle would have for them, given their situation, rather than for ourselves. While applying the contract test does not require one to explicitly go through the above steps, it does therefore require a capacity for perspective-taking.

Psychologists tend to use the term perspective-taking to refer to the capacity to consider another particular person's thoughts or feelings regarding something. It is often described to be either the capacity to imagine yourself in another's situation, or the capacity to imagine being the other in his situation (e.g. Batson, Early, & Salvarani, 1997a; Gordon, 1995). Applying a contract test requires a somewhat different type of perspective-taking, however. I shall now discuss two ways in which the perspective-taking that needs to occur in the contract test differs from this, drawing again on Scanlon's contract theory. I will then also explain why it yet counts as perspective-taking.

The first difference concerns what we are trying to find out when considering other standpoints. In order to find out whether an action is permissible, we need to find out not whether others would in fact agree to principles that permit it but whether *they have reason* to do so. We are thus not first and foremost interested in what they *would* think or feel about a principle.

We are, however, interested in a closely related question. As we need to determine whether others have reason to agree to a principle, we are interested in *what they would reasonably think or feel* with regard to the principle. We need to know what reasonable judgments they may have with regard to the principle; whether they could with reason object to the principle in question. This means we still need to consider the principle from alternative points of view than our own present point of view.

This can be clarified by considering Scanlon's contract test in somewhat more detail. We may distinguish between two steps that applying the test involves. The first step is to identify points of view affected by the principle under consideration. This requires us to think about the various ways in which a principle may affect others. We need to assess how the principle may affect

their liberty, opportunities, relations, self-respect, et cetera. Put more generally, we need to determine what costs and benefits acceptance of the principle would accrue to them. This obviously depends on properties associated with them and their situation. It may depend on their talents and capabilities, needs and vulnerabilities, on the social and economic conditions in which they are placed, on their cultural norms and and religious beliefs, and on their projects, aims, and tastes. Once aware of how they are affected given their position with regard to the principle, the second step is to determine whether they have, given these implications, reason to object to the principle. Determining whether others have reason to agree to a principle thus requires an understanding of their position with regard to the principle, even if we do not need to determine what they would in fact think or feel about it.[15] Note again that the above is a rational reconstruction: persons do not need to explicitly go through these steps to see whether or not others have reason to agree or to disagree with a given action or principle.

The second difference between the perspective-taking that psychologists are interested in and the perspective-taking required for applying the contract test concerns the sort of perspectives that are considered. Psychologists tend to be interested in our capacity to consider the perspectives of particular others. Contract theorists, on the other hand, tend to concentrate on abstract viewpoints. As Scanlon writes, while "we naturally think first of the specific individuals who are affected by specific actions" (p. 202), when applying the contract test "we must take a broader and more abstract perspective" (p. 202). Contract theorists hold that principles are justified if everyone has reason to accept them. But we can hardly consider every individual's viewpoint. That is one reason why contract theorists tend to concentrate on a more limited set of representatives. As Scanlon writes:

> in deciding which systems of principles are "acceptable", we cannot envisage the reactions of every actual person. We can consider only representative cases, and take into account only those objections that a person could raise while recognizing the force of similar objections by others." (Scanlon 1998, p. 171)

Scanlon calls such representative cases 'generic standpoints'. We must consider the standpoint of 'an agent' who would be required to act in a certain way, the standpoint of 'a person' who would be affected by such actions, the standpoint of 'a bystander', and so forth. Or as Scanlon puts it in an exchange with Parfit, "what we consider are not the reasons of actual persons but the

[15] Of course, when we are considering whether a given *reasonable* person has reason to object to a principle, considering how he would think or feel about the principle should be informative.

'generic' reasons that someone would have in virtue of occupying a certain role in regard to the principle in question" (2011, p. 131).[16]

Considering a generic standpoint is clearly different from considering a particular person's point of view. Generic standpoints are not bound to specific individuals, but may be occupied by groups of indeterminate persons united by virtue of certain properties. But the difference between these modes of perspective-taking seems to be a matter of degree rather than kind. One uses the same capacity in a somewhat different way. Compare considering the principle that allows the breaking of promises from the perspective of a specific person to whom a promise would be broken to considering that principle from the generic standpoint of *any* person to whom this would happen. In both cases, doing so would first and foremost involve an estimation of the implications for a person when a promise to him were to be broken. The main difference seems to be that to understand a specific person's point of view one may have to take into account additional idiosyncratic properties that would affect it.[17]

Furthermore, *identifying* relevant generic standpoints may require one to first consider the points of view of actual persons. As we just saw, generic standpoints that we must take into account are defined in terms of the reasons that individuals occupying them have. To occupy a generic standpoint with regard to a principle is to have reasons, in virtue of one's situation and other general characteristics, for wanting a principle to be rejected or accepted. Identifying relevant standpoints thus requires knowledge of the implications of a principle for others. Without such knowledge one would simply not know which standpoints to take into account. Take for example the generic standpoint of a person in extreme poverty who would be negatively affected if people were to stop donating to organisations such as Oxfam. The natural way

---

[16] It may be worth noting that Rawls (1999) defends a similar view: "In applying the two principles of justice to the basic structure of society one takes the position of certain representative individuals and considers how the social system looks to them. The perspective of those in these situations defines a suitably general point of view. But certainly not all social positions are relevant. For not only are there farmers, say, but dairy farmers, wheat farmers, farmers working on large tracts of land, and so on for other occupations and groups indefinitely. We cannot have a coherent and manageable theory if we must take such a multiplicity of positions into account. The assessment of so many competing claims is impossible. Therefore we need to identify certain positions as more basic than others and as providing an appropriate standpoint for judging the social system" (pp. 81-82).

[17] It is worth mentioning that once a property associated with a person's point of view affects that person's reasons for objecting to a principle, it becomes a relevant aspect of his generic standpoint.

to recognise such a generic standpoint is to consider the perspectives of persons who are actually in such a situation.

Not all relevant generic standpoints need to be identified through first considering the perspectives of particular others, though. Many moral principles have implications that a person applying the contract test may also experience herself. These principles thus include generic standpoints that she herself may also occupy with regard to such principles. Say, again, that I consider a principle that permits people to break promises for the sole reason of not feeling like keeping the promise. Relevant standpoints for evaluating whether this principle can reasonably be rejected include the standpoint of an agent to whom the promise would be broken and the standpoint of an agent who would otherwise be required to keep his promise even though he does not feel like it. Assuming that I am not presently in either of these two situations, I must imaginatively 'put' myself in them when considering them. But as these are situations in which I would be regularly put myself, I do not need to consider the perspectives of specific others to identify them.

However, many moral principles also have implications for certain persons that a person applying the contract test would *not* experience herself if the principles were generally accepted. To clarify this, I should first note that a generic standpoint may be quite specific. A given principle may affect different individuals in very different ways, depending on their capabilities, aims, and the conditions in which they are placed, among other things. Scanlon gives the example of a principle that forbids persons to break agreements. Such a principle has different implications for most of us than for persons who are, as Scanlon puts it, less able to "foresee possible difficulties and to resist subtle pressures to enter an agreement" (p. 205). As such persons are more easily drawn into unwelcome agreements, they have a relevant different point of view. For another example, consider a principle that prohibits abortion. With regard to such a principle, women occupy a different standpoint than men because they would be differently affected; similarly, individuals pregnant of a healthy foetus occupy a different standpoint from individuals with a foetus that would develop into a severely disabled child; an individual who is pregnant as the result of rape occupies a different standpoint than a person who became pregnant voluntarily; and there are surely additional relevant standpoints to distinguish. As generic standpoints may be quite specific, identifying them may require detailed information about others and their situations.

I conclude that even though there are some differences with the perspective-taking that psychologists are typically concerned with, applying Scanlon's contract test requires a variant of the perspective-taking they are interested in. Determining whether a principle is acceptable for others requires one to consider it from various standpoints different from one's present point of view. This conclusion can be generalised to other contract theories. As we saw, every contract test requires one to consider which principles persons who are differently situated than oneself, and whom therefore have different points of view regarding the principles, would agree to. I shall now briefly explain this in somewhat more detail with regard to Gauthier's contract theory which is, we saw in the previous chapter, rather different from Scanlon's.[18]

As I explained in the previous chapter, Gauthier identifies morality with those arrangements that would be the object of agreement among representatives of ours who would be bargaining about their terms of interaction. To identify whether a given arrangement would be among them, one must consider the perspectives of these representatives with regard to that arrangement and determine whether each of them has, given their particular situation, reason to accept it. Scanlon and Gauthier diverge with regard to the question of what it means for a person to have reason to accept an arrangement. For Scanlon, we saw, a person may only reject an arrangement if its acceptance burdens him more than any alternative arrangement. To what extent the arrangement burdens him depends on objective aspects of his situation, not on his subjective preferences. Gauthier holds a very different position. On his view, whether a bargainer has reason to accept an arrangement *does* depend on the extent to which it satisfies his preferences in comparison with alternative arrangements. Applying Gauthier's contract test thus requires one to gather different and, it seems, additional information to what is required for Scanlon's contract test. The contract tests both crucially involve perspective-taking, however: in both cases, one must consider principles from other points of view than one's own.

---

[18] It may be worthwhile to note this is also the case for Rawls. In order to follow Rawls's argument that the two principles of justice would be chosen from the original position, one must engage in perspective-taking at two different points. First, one is to consider the standpoint of a person who does not know any particular facts about himself. While Rawls may not require us to fully imagine ourselves behind a veil of ignorance, it does require us to attain an understanding of this rather peculiar perspective. Second, from behind the veil of ignorance, one is to consider what various distributive principles would mean for representative social positions that one may, for all one knows, come to occupy. Again, this requires one to achieve an understanding of perspectives different from one's own.

This section argued that perspective-taking is crucial for applying a contract test. But it is of course not the only capacity required for doing so. Applying the contract test requires other general capacities, such as a capacity to predict consequences, a capacity to subsume individual actions under more general principles, and a capacity to recognise and weigh reasons. A complete investigation of the Practicability Assumption requires also an investigation into these capacities. Perspective-taking, however, has a particularly central role in the contract test. And in contrast to the other general capacities required for applying the contract test, it does not seem to have as central a role in the justificatory procedures proposed by other moral theories.[19] It is therefore an interesting capacity to focus on when studying the empirical plausibility of contract theory. What may be more important, there is some reason to worry whether our capacity for perspective-taking is up to the task.

5 How to evaluate the Practicability Assumption

In the preceding sections I have argued that besides being of interest for agents who are motivated to use the contract test, an investigation into our ability to use the contract test is relevant for an evaluation of contract theory itself. Contract theories are committed to a variant of the Practicability Assumption. In particular, I have argued that moral contract theorists such as Scanlon and Gauthier assume that actual agents can, without too much difficulty, learn to apply the contract test adequately under circumstances that include those typical of everyday life.

In the following chapters I shall investigate the plausibility of this assumption in the light of empirical findings. In particular, I shall consider what implications findings on our capacity for perspective-taking have for the Practicability Assumption. The present section has two related goals. The first is to describe in what way the Practicability Assumption will be investigated in the following chapters. The second is to introduce some concerns about the plausibility of the Practicability Assumption, which emphasise the relevance of such an investigation.

What does it mean for the Practicability Assumption to be plausible? An assumption is plausible, I shall assume, if there are considerations counting in favour of it being satisfied, and that these outweigh considerations that count against it. When the assumption concerns empirical facts, considerations

[19] That is not to exclude that they are nevertheless committed to assumptions about perspective-taking.

counting in favour of it come in the form of supporting empirical findings, and considerations counting against it as opposing evidence. Plausibility comes in degrees, and we can thus say that the plausibility of an empirical assumption depends both on the strength of supporting findings and on that of opposing findings. Opposing findings tend to weigh heavier against an assumption than supporting findings can weigh in favour of it; opposing findings can in principle reveal an assumption to be false, whereas supporting finding can only make it more likely to be true.

No psychologist has taken up the task to experimentally investigate our ability to apply particular contract tests. However, the last few decades have seen much research in social cognitive abilities such as the ability for perspective-taking. As I explained in the previous section, applying a contract test requires considering principles for the general regulation of behaviour from different perspectives than one's own. The Practicability Assumption thus requires that our capacity for perspective-taking is up to the task. The first part of my investigation will concentrate on the question of whether this is the case: whether the Practicability Assumption is plausible in the light of existing findings on perspective-taking.

This requirement about perspective-taking can be made more precise by separating it into two different requirements. First, in order for people to be able to apply the contract test to evaluate actions or principles, it must be the case that they are able to arrive at moral judgments about such objects on the basis of reasoning that involves perspective-taking. If they are not able to do this, they will not be able to apply the contract test under those circumstances. Second, in order for people to be able to apply the contract test *adequately*, which means that they can use it reliably and get determinate results, it must be the case that their perspective-taking is sufficiently accurate. Without being able to identify and understand relevant perspectives, they are likely to either fail to find out whether a principle would be the object of agreement or not, or reach a mistaken conclusion regarding this.

I will evaluate the plausibility of the Practicability Assumption by investigating whether empirical findings suggest that these two requirements are met or not. Although their being met would not imply that the Practicability Assumption itself is correct, it would enhance its plausibility. I will now explain both of these requirements separately in terms of how they can be investigated. More precisely, I will explain what kind of findings would support them or oppose them. I will also introduce for each of them a concern: a consideration for thinking that they may not be met.

The first requirement will be the subject of Chapter 3. I will examine whether persons *can* form moral judgments by a reasoning process that involves perspective-taking by considering whether empirical findings suggest they *do* sometimes form judgments through such a process. If it turns out that their judgments are under certain circumstances arrived at through reasoning that involves perspective-taking, it follows that this can be done in that way. Findings on the role of perspective-taking in moral judgment can thus provide support for this first assumption. On the other hand, if we find that perspective-taking does not inform moral judgment under those circumstances, we would also have no reason to believe that we can form judgments through a procedure such as the contract test.

Findings on perspective-taking may also provide evidence which *undermines* this first requirement. This may seem unlikely. From the fact that persons do not in fact arrive at moral judgments in a particular way under designated circumstances it does not follow that they *cannot* do so. Inability is just one possible explanation for why persons do not form moral judgments in this way; it could also be explained by their not exercising the ability for other reasons. Studies on perspective-taking and its relation to moral judgment can, however, provide support for the inability explanation. For example, were we to find that people do not form moral judgments through perspective-taking when they are under time pressure, and in addition find that perspective-taking is an effortful and time-consuming process, this additional finding would support thinking that people cannot form moral judgments through perspective-taking when cognitive resources are low. By considering the limitations and constraints on perspective-taking we could thus find evidence that undermines this first requirement.

As I mentioned before, there is in fact reason to be concerned that the first requirement may not be met. Several moral psychologists have recently argued that moral judgments are typically based not in reasoning, with or without perspective-taking, but in intuitions or emotions (e.g. Haidt, 2001; Nichols, 2004; Prinz, 2007). These intuitions are taken to be the result of quick, unconscious, effortless processes. If this account is correct, the concern arises that we may not be able to use the contract test as a moral guide. If we normally arrive at our moral judgments in a very different way than on the basis of reasoning, we may have difficulty in forming judgments through a reasoning procedure such as the contract test. Not only is our cognitive plasticity limited, there may be a good reason for why we usually would not form judgments on the basis of reasoning. For example, we may not have

sufficient cognitive resources for doing so. By examining studies regarding the role of perspective-taking in moral judgment, Chapter 3 will assess to what extent empirical findings substantiate this concern.

Chapter 4 considers the second requirement. To assess whether perspective-taking can inform moral judgment accurately we would ideally draw on studies in which our perspective-taking accuracy is investigated in a moral context. Unfortunately, there are not many such studies available. However, there is a larger number of studies regarding the accuracy of perspective-taking in contexts that are not specifically moral. I shall in general assume that findings from such contexts can be generalised to situations in which we would consider other points of view for arriving at moral judgments.

As with the first requirement, there is reason to be concerned that this second requirement may not be met. Empirical studies find that persons are not very accurate perspective-takers. They often fail to see that others have different viewpoints from their own. And even when they do recognise this, they still tend to overestimate the similarity with their own present perspective. Social psychologists have argued that such perspective-taking inaccuracy occurs because we base our interpretations of other perspectives on our own present point of view (Epley, 2008; Epley & Caruso, 2004). Moreover, they have suggested we may often be too stuck in our own point of view to be able to overcome such inaccuracies. The concern presented here is thus not only that our perspective-taking accuracy tends to be limited, but that there may be little we can do about it.

As Chapter 3 and 4 concentrate on our present ability for perspective-taking, they provide first and foremost information about how able persons are *presently* at applying the contract test. But as I explained before (§3.3), the Practicability Assumption requires that we can *come* to apply the contract test. Chapter 5 will therefore focus on the question of to what extent persons can overcome the limitations associated with our capacity for perspective-taking.

6 Conclusions

This chapter put forward two central claims. The first is that in proposing the contract test as a procedure for moral justification, moral contract theories are committed to the assumption that actual agents can, without too much difficulty, learn to apply the contract test adequately under circumstances that include those typical of everyday life. Investigating how able we are at

applying the contract test is therefore not only relevant for agents motivated to use it, but also for an assessment of the plausibility of contract theory itself.

The second central claim of this chapter is that findings on perspective-taking are of relevance for the Practicability Assumption. Determining whether a principle would be the object of agreement requires an agent to consider it from other standpoints than just her own present point of view. Applying the contract test adequately thus requires her perspective-taking capacity to be up to the task. The following chapters will consider what findings on perspective-taking reveal about the Practicability Assumption's plausibility.

3

Perspective-Taking in Moral Judgment

1 Introduction

Contract theorists give perspective-taking a central role in their procedures for moral justification, I have argued in the previous chapter. To test whether an action is justified one must consider whether principles that permit it are acceptable from points of view different than one's own. I will now investigate whether this contract test fits with how people actually form moral judgments. The reason for doing so is to examine what I called the Practicability Assumption: that actual persons can come to apply the contract test adequately in circumstances that include those of everyday life. This assumption must be satisfied for persons to be able to rely on the contract test as a moral guide. I have argued that moral contract theorists such as Scanlon and Gauthier are committed to this assumption.

The previous chapter (§5) introduced two requirements associated with the Practicability Assumption: that we are, under the appropriate circumstances, able to apply a reasoning procedure that involves perspective-taking such as the contract test, and that our accuracy in perspective-taking is sufficient for us to apply the test adequately. The present chapter is concerned with the first requirement. The most straightforward way of evaluating its plausibility is by examining whether this method of forming moral judgments fits with how persons actually form moral judgments.

That our moral judgments are based on reasoning that involves perspective-taking has been a dominant position in the field of moral psychology. Both Jean Piaget and Lawrence Kohlberg, the most influential moral psychologists of the 20th century, held that mature moral judgment is based in reasoning that involves adopting perspectives different from one's

own. The position has also been popular among moral philosophers.[1] In the last few decades, however, this view has received much criticism. According to some, such *perspective-taking accounts of moral judgment*, as Shaun Nichols (2004) has called them, present a far too intellectual view of moral judgment that does not fit recent empirical findings. Along with several other moral psychologists, Nichols has rejected the idea that reasoning with perspective-taking has a significant role in moral judgment in favour of the idea that moral judgments are typically based in intuitions or emotions (e.g. Haidt, 2001; Prinz, 2007). As I mentioned before (2§5), in contrast to reasoning, these intuitions are taken to be the result of quick, unconscious, effortless processes.

It would not be good news for the Practicability Assumption if these critics were correct that neither reasoning nor perspective-taking have a significant role in moral judgment. It supports the concern that we may not be able to come to rely on the contract test as a moral guide. As our cognitive plasticity is limited, it may be difficult or even impossible to adopt a method of moral judgment that is alien to how we normally form moral judgments. In order to find out whether this concern is serious, the present chapter shall examine to what extent the criticism of perspective-taking accounts is supported by empirical findings.

I will also examine whether there is evidence in favour of the idea that people form moral judgments on the basis of reasoning that involves perspective-taking. Reasoning refers here to *practical* reasoning, the process of assessing and weighing reasons for action (Wallace, 2008). As I explained before, to engage in perspective-taking is to consider events from an alternative point of view than one's own. By studying whether people form moral judgments through reasoning that involves perspective-taking we can learn something about the plausibility of the Practicability Assumption. If people in fact form moral judgments in this way, they certainly *can* do so. Of course, a contract test is a specific type of reasoning procedure that involves a

[1] Thomas Nagel, John Deigh, and Stephen Darwall are three examples. Nagel (1970) writes that the principle of altruism "arises from the capacity to view oneself simultaneously as 'I' and as someone—an impersonally specifiable individual" (p. 19) Deigh (1995) argues that in order to make what he calls "deeper judgments of right and wrong" one must have a capacity for "taking another's perspective [so that one] sees the purposes that give extension and structure to the other's life and sees those purposes as worthwhile, as purposes that matter" (p. 760). Darwall (2006) claims that in order to recognize obligations that we have towards others we must adopt what he calls the second-person standpoint, which "requires empathy or the capacity to put oneself in another's shoes" (p. 44). As I mentioned in the previous chapter, It is also adopted by Scanlon when he suggests his contractualist account of moral reasoning fits with how we actually from moral judgments.

specific kind of perspective-taking. For that reason, the finding that people engage in *some* process of reasoning that involves perspective-taking as such does not yet prove that people can indeed apply the contract test. But it does make it more plausible.

It is worth emphasising that for there to be reason to believe that people can form moral judgments through a process of reasoning that involves considering perspectives different from one's own—or, as I will usually write, form moral judgments through perspective-taking—it need not be found that all or even most of our judgments are derived this way. It is sufficient if persons sometimes form moral judgments through perspective-taking in the appropriate circumstances. However, the more central a role this sort of reasoning has in our actual moral judgment, the more plausible it is that we can come to adopt the contract test as a moral guide.

The structure is as follows. The next section examines whether empirical findings substantiate two challenges that have been posed against perspective-taking accounts of moral judgment. It considers whether there is evidence against the idea that persons form moral judgments through reasoning that involves perspective-taking. After that, the third section examines what evidence there is in favour of it.

2 Two challenges against perspective-taking accounts of moral judgment

The view that people form moral judgments through reasoning that involves perspective-taking used to be dominant in psychology. Such a view, if correct, enhances the plausibility of the Practicability Assumption: the contract test would fit with how we normally arrive at moral judgment, such that adopting it as a moral guide should not be too difficult. The view has, however, been challenged in recent years. Empirical findings on the moral judgment of individuals with autism and children have been argued to reveal that perspective-taking has no significant role in moral judgment. Moreover, it has been argued that moral judgments are typically caused by intuitions and emotions rather than reasoning. Such challenges to perspective-taking accounts of moral judgment not only call into doubt a potential source of support for the Practicability Assumption, but also, as I mentioned above, fuels the concern that persons may not be able to adopt the contract test as a moral guide. I shall therefore examine in this section to what extent they are backed up by empirical findings.

2.1 First challenge: moral judgments are typically caused by intuitions rather than reasoning

It has become a common view in moral psychology that moral judgments are often not the result of occurrent reasoning but of intuitions or emotions (as was mentioned in 2§5). Research on 'dumbfounding' is a case in point. Haidt and colleagues (2000) found that participants are sometimes unable to justify their moral judgments. In their study they asked participants to express moral judgments about scenarios in which protagonists engaged in behaviours that under many circumstances would be considered impermissible. For example, they were asked to judge about a scenario that involved a brother and sister making love to one another. After expressing their judgments, participants were asked to provide a justification. Although participants usually showed little difficulty in coming up with reasons, they often referred to considerations that were explicitly excluded in the scenarios at hand. With regard to the aforementioned incest scenario, for example, it was stipulated by the researchers that the siblings were adults, that the sex was fully consensual, that birth control was being used, and that there were no negative emotional consequences; furthermore, participants were told that both brother and sister enjoyed the experience, cherished it, and as a consequence grew even closer. That did not stop participants from invoking common objections against incest, however.

Most interestingly, after having been made aware of the fact that given considerations did not apply to the scenario at hand, many participants nonetheless stuck to their judgment, expressing thoughts such as "I know it's wrong, but I just can't come up with a reason why" (Haidt et al., 2000). Haidt (2001) has argued that findings such as these suggest that moral judgments are often not caused by reasoning, but by intuitions: "the sudden appearance in consciousness of a moral judgment, including an affective valence (good–bad, like–dislike), without any conscious awareness of having gone through the steps of searching, weighing evidence, or inferring a conclusion" (Haidt, 2001, p. 818). Rather than thinking about the details of the situation before forming a judgment about two siblings having sex, we judge it is wrong because it 'feels' wrong. As Haidt shows, this conclusion fits well with other studies suggesting that moral judgments may arise through unintentional and uncontrollable psychological processes that operate outside of conscious awareness (Epley & Caruso, 2004; for overviews see Haidt, 2001; 2007).

I take findings such as these to suggest that not each and every moral judgment that we have is caused by occurrent reasoning that involves perspective-taking. Indeed, it may well be true, as Haidt believes, that most of our moral judgments are caused by intuitions rather than by such occurrent reasoning. Besides putting pressure on perspective-taking accounts of moral judgment, this view also has implications for the Practicability Assumption. Of course, from the fact that we often do not form moral judgments through reasoning that involves perspective-taking it does not, by itself, follow that we cannot. There is however a plausible explanation for why people would rely on intuitions rather than reasoning that does support this conclusion. Often when we form moral judgments we are constrained by time and the distractions of everyday life. Our cognitive resources are relatively low. Whereas intuitions can arise rapidly, unintentionally, and effortlessly, reasoning requires conscious attention and effort. There is thus reason to be concerned that we sometimes cannot form moral judgments on the basis of reasoning that involves perspective-taking.

This is supported by the fact that, besides reasoning, perspective-taking also requires cognitive resources. Several studies have found that when people are under time pressure or distracted by another task, they fail to understand viewpoints different from their own (Davis, Conklin, Smith, & Luce, 1996; Epley, Keysar, Van Boven, & Gilovich, 2004; Lin, Keysar, & Epley, 2010). For example, participants were told about Tom who, during dinner with his friends Steve and Gina, was urged by Gina to see a comedian who according to her was "just hilarious". Some participants were then told that Tom went to see the comedian and that he loved it, while others were told that he hated it. After this, participants heard a voicemail message in which Tom says to Steve that "all I can say is that you have to see him yourself to believe how hilarious he really is". They were then asked to judge whether the message was sincere or sarcastic. The researchers found that participants who had been told that Tom hated the comedian's show were more likely to think an uninformed listener would interpret the message as sarcastic than participants who had been told he enjoyed it. Apparently, participants did not adjust sufficiently for their privileged information—an instance of what Epley and colleagues call the egocentric bias.

Most interesting for our purposes, this bias became substantially stronger when participants had to give their answer under time pressure. When participants who had been told Tom hated the comedian had to predict Steve's interpretation of Tom's ambiguous voicemail message within 3 seconds after

hearing it, they interpreted it sarcastically 66% of the time, whereas participants who had been asked to consider their judgments carefully and respond at their leisure did so only 50% of the time. Apparently, the researchers conclude, perspective-taking requires an amount of effort that is not always available.

If doing such a relatively simple perspective-taking task is already difficult under low cognitive resources, than applying the contract test certainly is. Evaluating a principle by means of the test typically requires considering it from multiple standpoints, which requires more cognitive resources than only considering one particular point of view. Findings such as these thus suggest it is unlikely that people can apply the contract test when their cognitive resources are low.

What would this conclusion mean for the Practicability Assumption? In the previous chapter I argued that, by the principle that ought-implies-can, moral contract theorists are committed to holding that persons can abide by moral principles that would be the object of agreement (2§3.2). I also argued that for persons to do this consistently, they must be able to assess whether potential actions they may perform conform with such principles. Persons must thus be able to test their actions by means of the contract test. That persons cannot apply the contract test when their cognitive resources are low implies that there are situations in which they cannot examine their actions by means of the contract test. It does not follow, however, that they cannot rely on the contract test as a moral guide.

The crucial idea, already alluded to in the previous chapter (§3.2), is that for the contract test to be a moral guide it does not need to always have a synchronic role in moral judgment but may also have a diachronic role.[2] When a given moral judgment is the direct result of applying the contract test here and now, the contract test has a *synchronic* role. As moral contract theorists such as Gauthier and Scanlon do not propose their respective contract tests as decision procedures by which persons must consciously make all their choices, they do not hold that all our judgments must be arrived at in this manner. The contract test is a device to derive moral principles that ought to be followed. Persons may be able to *internalise* such principles and subsequently follow them in situations they apply to. The contract test has in that case not a

[2] I borrow the distinction from Jesse Prinz (2011).

synchronic but a *diachronic* role.[3] This provides a potential solution to the problem of limited cognitive resources: contract theorists may propose that persons use the contract test diachronically to prepare themselves for future situations in which they have insufficient resources to apply it (or use it synchronically).

This proposal may seem to contradict something I said in the previous chapter. I there rejected the idea that persons can ensure they act in conformity with principles that satisfy the contract test by preparing themselves from a cool hour. Drawing on Scanlon, I argued that given that such principles must be succinct and general in order to be rememberable they will not cover the complexity and variance associated with everyday situations.

The important difference between that proposal and the present proposal, however, is that the present proposal does not confine the learning of moral principles to a cool hour. Persons may internalise moral principles through actual moral practice. Take for example the familiar moral principle that promises made freely should be kept. As Scanlon points out, the principle is far more complex than a statable rule:

> Anyone who understands the point of promising—what it is supposed to ensure and what it is to protect us against—will see that certain reasons for going back on a promise could not be allowed without rendering promises pointless, while other exceptions must be allowed if the practice is not to be unbearably costly. [...] All of this structure and more is part of what each of us knows if we understand the principle that promises ought to be kept. In making particular judgments of right and wrong we are drawing on this complex understanding, rather than applying a statable rule, and this understanding enables us to arrive at conclusions about new and difficult cases, which no rule would cover. (Scanlon, 1998, pp. 200-201)

The moral understanding that Scanlon is talking about cannot be acquired by applying the contract test in a cool hour. But persons may acquire it in everyday moral practice. We are continuously confronted with situations in which promises may or may not be kept. In such situations, we may often have some time to examine whether keeping or not keeping the promise would be in conformity with principles that everyone has reason to agree to or not. That is to say, we may be able to apply the contract test. By doing so, we may develop the moral understanding regarding promises Scanlon is talking about.

---

[3] In a similar vein, a defender of a perspective-taking account of moral judgment can respond that reasoning involving perspective-taking has a diachronic role. I shall say more about this below.

We would gradually internalise a complex moral principle that would be the object of agreement.

Internalising moral principles may make it easier for us to arrive at moral judgments when cognitive resources are low. There is good reason to think that persons may through reasoning change or 'educate', as Hanno Sauer calls it, their moral intuitions (Fine, 2006; Pizarro & Bloom, 2003; Sauer, 2012). For example, it has been found that individuals who decide to become vegetarians on the basis of moral considerations come to experience emotions of disgust with regard to meat that individuals who do not eat meat for health reasons and meat eaters do not have (Fessler, Arguello, Mekdara, & Macias, 2003; Rozin, Markwith, & Stoess, 1997).

There are some caveats to the proposal. First of all, given the large variety of moral scenarios, persons will time and again find themselves confronted with cases that are either new or difficult. In some of these cases, they may not have sufficient cognitive resources to apply the contract test. Consequently, even if the proposal would work, it would not fully solve the problem of limited cognitive resources.

Moreover, the proposal does require that certain empirical conditions are met. First, it requires that the contract test can be used synchronically sufficiently often in moral practice. To internalise principles that everyone has reason to agree to in such a way that we come prepared to situations in which our cognitive resources are low, persons must be able to apply the contract test in similar situations. Second, the proposal requires that the contract test can indeed be used diachronically: that persons can internalise moral principles by applying it and draw on these when their cognitive resources are low. The proposal thus leads to a slightly modified first requirement of the Practicability Assumption: that a reasoning procedure that involves perspective-taking such as the contract test can have both a synchronic and a diachronic role in moral judgment. Again, the most straightforward way of evaluating this requirement is by considering whether reasoning that involves perspective-taking has in fact these roles in moral judgment. I shall thus continue investigating evidence for and against the view that reasoning that involves perspective-taking has a significant role in moral judgment.

2.2 Second challenge: perspective-taking has no significant role in moral judgment

Already early in development children appear to be able to make moral judgments. One finding that in particular has received much attention is that

even young children can distinguish between conventional and moral norms. Children tend to assign a different status to rules such as 'do not pull another child's hair' than conventions such as 'do not eat gum in class'. Smetana and Braeges (1990) found that children of almost three years old were more likely to judge that moral violations generalized across contexts than conventional violations when asked "At another school, is it OK (or not OK) to X?". According to Shaun Nichols, this reveals that children of this age have already developed the essentials of moral judgment—what he calls 'core moral judgment' (Nichols, 2004).

Nichols (2004) argues that this finding does not sit well with the idea that we form moral judgments through reasoning that involves perspective-taking. A perspective-taking account would have to explain the fact that a child judges pulling another child's hair as less permissible than eating gum in class in terms of the child's capacity for perspective-taking. Importantly, in line with the proposal from the previous section, such an account may ascribe a diachronic role rather than a synchronic role to perspective-taking: it may state that reasoning that involves perspective-taking lies at the basis of the understanding of right and wrong and associated moral concepts that is expressed in moral judgments. Although such an account would not be committed to the claim that whenever a child judges a particular transgression to be impermissible it engages in perspective-taking, Nichols says, it is committed to claim that the child has gone through such a process at some point in time, at which she came to believe that type of transgression as being impermissible.

Nichols (2004) claims such an explanation of the child's moral understanding is implausible. Perspective-taking is an advanced social cognitive capacity that is not fully developed in the three-year-old. Indeed, it is widely believed that children below the age of three have a rather limited understanding of the minds of others. This is mostly based on the well-known 'false-belief task'. In the classic version of this task, children watch a puppet show in which one puppet, Maxi, puts chocolate in a box before going out to play. While Maxi is out, his mother comes in and moves the chocolate from the box to the cupboard. Children are then asked where Maxi will look. While children beyond the age of four tend to say that Maxi will look in the box, thereby displaying that they take into account that Maxi has a different perspective on the situation than they themselves have, children before the age of three tend to say that Maxi will look in the cupboard rather than in the box where he stored it. Nichols concludes from this that perspective-taking is

unlikely to explain children's moral judgments. If their capacity to understand perspectives different from their own is limited, it is unlikely to be involved in their understanding of the difference between conventional and moral rules, Nichols argues. He concludes from this that core moral judgment does not depend on perspective-taking.

Nichols (2004) makes a similar argument on the basis of findings on the moral judgment of individuals with autism. Anecdotal evidence as well as experimental evidence shows that individuals with autism can develop moral judgment. For one, they have been found to distinguish between moral and conventional rules (James & Blair, 1996). They are, however, limited in their capacity for perspective-taking. Indeed, impairments in one's capacity to understand others is one of the defining characteristics of autism as well as other disorders in the so-called autism spectrum, such as Asperger's Syndrome. Given this impairment, it is not very plausible that their moral understanding has a basis in perspective-taking.

On the basis of such considerations, Nichols writes that "data from development and psychopathology pose an obstacle for the perspective-taking account of core moral judgment" (Nichols, 2004, p. 11). He takes it to indicate that the capacity for core moral judgment and the capacity for perspective-taking are dissociated.[4] If one can have moral judgment without perspective-taking, perspective-taking is apparently not involved in basic moral judgment. While Nichols does not explicitly claim that perspective-taking does not have any role in moral judgment—he writes at one point that "[s]urely people sometimes use perspective taking in moral evaluations" (p. 9)—he goes on to develop a sentimentalist theory of moral judgment in which it effectively has no significant role. If this theory were correct, perspective-taking would have at best a peripheral role in our moral thinking. It would certainly not have a sufficiently significant role to provide us with reason to think that persons can, without too much difficulty, adopt the contract test as a moral guide.

There are two crucial premises in Nichols's argument against the idea that perspective-taking has a significant role in moral judgment. The first is that the grasp that children and individuals with autism have of perspectives different from their own is so limited for children that it cannot explain their

[4] Nichols (2004) in fact argues for a double dissociation between moral judgment and perspective-taking: that one can have moral judgment and no perspective-taking, as well as the other way around. He uses empirical studies on psychopathy to defend this latter point. Given that the psychopath has a capacity for perspective-taking but not a capacity for moral judgment would at most show that perspective-taking is not sufficient for moral judgment, I do not discuss it here.

moral understanding. The second is that they have core moral judgment; that they have reached a mature level of moral understanding, not too different from that of adults. Without this second premise, the premise that perspective-taking is not involved in the moral judgment of children could not imply that perspective-taking does not have a central role in the mature moral judgment in which we are mostly interested. I shall consider to what extent these two premises are supported, starting with the first.

Compared to normal adults, children as well as individuals with autism have a restricted grasp of perspectives different from their own. But how restricted exactly? Let's start with children. As I mentioned before, it has been found that most children under the age of four fail false-belief tasks. This has long been thought to indicate that they have a rather limited understanding of other minds. But there are other explanations for this failure. The false-belief task does not only require perspective-taking, but also additional capacities that are not as well developed in the average three-year-old as in the average four-year-old. Several studies have shown that when the task is made cognitively less complex in certain ways, the performance of three-year olds improves. When children are given more time to absorb required information, they are twice as likely to pass the test (Zaitchik, 1991). This reveals the importance of general information-processing capacities in the false-belief task. Furthermore, when three-year-olds are only told of the actual location rather than seeing it themselves, they tend to pass the test (Zaitchik, 1991). This suggests that the explanation that three-year-olds often fail the false-belief test is not that they cannot grasp other points of view but that they have difficulty inhibiting salient information (Goldman, 2006). This is supported by several experiments that have found that three-year olds tend to make more egocentric mistakes than somewhat older children when judging the desires or beliefs of others while having contrary desires and beliefs themselves, and much less when they do not have such mental states (Birch & Bloom, 2003; Moore et al., 1995). Birch and Bloom (2003) conclude that such findings suggest children's difficulty in thinking about false-beliefs "is better characterized by the curse of knowledge than by more general egocentrism" (p. 283).

That the failure of three-year-olds in passing the false-belief task does not imply that they have no understanding of other points of view is perhaps most convincingly suggested by studies that rely on an implicit measure of understanding false beliefs. It has been observed that whereas three-year-olds tend to give a false verbal answer, they turn out to look in the right place

(Clements & Perner, 1994). Using looking behaviour as a measure, Onishi and Baillargeon (2005) report that even 15-month-olds have some understanding of false-beliefs. Infants were observed in two conditions: a condition in which they could see that a person who just placed an object in a box watched as it was being moved, and a condition in which the infants could see that the person could *not* see the object being moved. The looking behaviour of infants suggested that in the first condition they expected the person to look at the place where the object actually was, whereas in the other condition they expected the person to look at the place where the object was originally placed. This suggests even 15-month-olds have a grasp of the perspectives of others.

That very young children can recognise alternative perspectives than their own is also confirmed by the fact that young children respond to the distress of others (Goldman, 2006). Newborns already respond to the emotions of others, for example when they start to cry upon hearing another baby crying. This phenomenon of emotional contagion does not require perspective-taking: the baby simply 'catches' the emotion of another baby. However, it does not take long before the child learns to distinguish another's distress from her own. Before the age of two, children demonstrate comforting behaviour in response to distress expressed by their mothers *without* necessarily becoming upset themselves (Zahn-Waxler, Radke-Yarrow, Wagner, & Chapman, 1992). Furthermore, children of this age often comfort in appropriate ways, displaying an understanding of the other's situation.[5] More directly related to our purposes, it has been found that they can use their understanding of another's perspective when making moral judgments. Helwig and colleagues (2001) found that 3-year-old children who make judgments about the wrongness of actions take into account others' psychological reactions, even when they are idiosyncratic and discrepant with their own.

Finally, that young children have some understanding of alternative perspectives is also confirmed by studies of pretend play. From 18 months onwards, children engage in pretend play, acting as if they were in a different situation than their actual one (Siegler et al., 2011). A few months later, children also engage in what the psychologist Paul Harris calls role play, imagining and acting out of the role of a person or creature (Harris, 2000).

[5] That is not to say that there's not yet much to learn. For example, when a young child sees a friend in distress she may fetch her own security blanket rather than the friend's blanket (Siegler, DeLoache, & Eisenberg, 2011).

Two-year-olds can give non-animate objects certain roles, and speak about them as if they experience the role from the viewpoint of the object. They may use deictic terms appropriate for the character, and give expression to its emotions, needs and sensations (Goldman, 2006). They can also recognize pretend play that others engage in, and share in it (Friedman & Leslie, 2007).

What about individuals with autism? In general, the social development of children with autism is much slower than that of normal children. Children with this disorder tend to show little responsiveness to the world in general, and appear not to recognise or care about those around them (Comer, 2003). Approximately half of them fail to speak or develop language skills, and those who do talk may show peculiarities in their speech. As such, many children with autism never reach the verbal intelligence of normal children. Amongst those who do, who are sometimes said to have 'high-functioning autism', social cognitive impairments remain. Such social cognitive impairments are also typical for so-called autism spectrum disorders, such as Asperger's Syndrome.[6] As mentioned before, one of the clearest findings is that autistic children who do manage to pass the false-belief test do so at a much later age than normal children do. They also have difficulty at other perspective-taking tasks, and show differences from normal children in pretend play (Hobson, Lee, & Hobson, 2008). Their pretend play starts later, is less frequent, tends not to occur spontaneously, and is limited (Jarrold, 2003). In addition, individuals with autism who are able to understand mental states of others when explicitly prompted to, typically fail to do so spontaneously from the observation of behaviour (Senju, Southgate, White, & Frith, 2009).

Although this all suggests that the capacity to understand other perspectives is impaired in individuals with autism, we should not be too quick to conclude that they have no capacity to do so altogether. Children with autism can engage in pretend play, even if they do it less spontaneously and less creatively (Hobson et al., 2008; Jarrold, 2003). They have also been found to adjust their language when talking to a naive instead of a knowledgeable listener (Geller, 1988). And although worse at it than normal children, they are able to do role-taking tasks that involve the telling and re-telling of stories from the perspectives of different protagonists, and even use mental states terms when doing so (García-Pérez, Hobson, & Lee, 2007). Moreover, many

[6] In the following, when I speak of individuals with autism I will be referring both to individuals who have high-functioning classical autism and individuals with Asperger's Syndrome. While there are important differences between these disorders, they are both characterized by a limited perspective-taking capacity, which is why they are of interest in the present study.

children with autism are eventually able to succeed in cognitive perspective-taking tests such as the false-belief task (Moran et al., 2011; Zalla & Leboyer, 2011).

In addition, individuals with autism are not necessarily impaired in capacities closely related to that of cognitively understanding another's point of view. They are not worse at adopting the visual perspective of others than normal subjects (David et al., 2009). They have been found not to be worse at recognizing emotions than children who have the same verbal skills (Castelli, 2005). Finally, there is some reason to think that individuals with autism can have affective empathy with others, even though affective responses may often fail to occur due to limitations in their understanding of the other (Blair, 2008). Taking this together, we should conclude that, while impaired, individuals with autism may have some grasp of perspectives different from their own.

So to what extent is Nichols's first premise warranted? Although young children do not have a fully developed capacity for perspective-taking, the above findings suggest they can grasp alternative perspectives. It is therefore far from clear that this capacity does not underlie whatever moral understanding they have. With regard to individuals with autism Nichols's first premise is more plausible. Both their capacity and their use of perspective-taking is very limited. There may thus be a better explanation of their moral understanding than what perspective-taking accounts can provide. One such explanation will be offered below.

As I mentioned before, there is a second crucial premise to Nichols's argument. Nichols claims that children and individuals with autism have 'core moral judgment'. This term suggests that their moral judgment is in essence comparable to the moral judgment of normal adults. Without this premise, the finding that children and individuals with autism have moral judgment without perspective-taking would be compatible with adults forming moral judgments through perspective-taking.

The three-year-old child shows remarkable moral understanding. Even so, it finds itself only at the beginning of moral development. There are significant differences between the moral judgment of children and that of adults. Most notably, young children judge actions mostly on the basis of their outcomes, and do not give intent as large a role as adults (Baird & Astington, 2004; Helwig et al., 2001; Zelazo, Helwig, & Lau, 1996). They are less likely to mitigate blame for accidental harms, and may judge accidental harms to be

worse than failed attempts to harm (Baird & Astington, 2004). This is stark contrast to adults, who tend not to judge accidental harm as blameworthy.

The moral *reasoning* of children is also less sophisticated than that of adults (Eisenberg, 1986; Rest, Narvaez, Thoma, & Bebeau, 2000; Siegler et al., 2011). As Barriga and colleagues (2009) write, there is "a well-documented worldwide age trend towards a deeper and more sophisticated understanding of moral decision and the justifications given for basic moral values and actions" (p. 255). The moral reasoning of young children is initially self-interested. Young children tend to justify moral judgments in terms of rewards and punishments. The developmental psychologist Nancy Eisenberg has called this mode of reasoning *hedonistic moral reasoning* (Eisenberg, Cumberland, Guthrie, Murphy, & Shepard, 2005; Siegler et al., 2011). There is abundant evidence that the child's reasoning becomes more and more advanced in the course of development[7].

Piaget and Kohlberg held that children's moral development occurs in a sequence of separate stages. While most developmental psychologists have let go of the idea that persons find themselves in one stage at a time, it is still commonly believed that there are separable schemas or modes of moral reasoning which follow a developmental sequence (Siegler et al., 2011).[8] These modes are associated with different levels of moral development. Hedonistic moral reasoning, which is the predominant mode for preschoolers and younger elementary school children, is associated with the first level. Reasoning in terms of one's affective relationships with others is also associated with this level (Lane, Wellman, Olson, LaBounty, & Kerr, 2010).

---

[7] Psychologists use tests to measure one's level of moral judgment. In the past, researchers made only use of clinical interview measures, such as Kohlberg's Moral Judgment Inventory. There are several disadvantages to such tests, such as that they are cumbersome to apply and laborious to train and learn to other experimenters. That is why researchers have turned to develop standardized paper-and-pencil tests. Two commonly used tests are the Defining Issues Test (DIT) and the Prosocial Reasoning Objective Measure (PROM). The DIT confronts participants with a series of moral dilemmas, each of which is accompanied with fragments of lines of reasoning about the dilemma. Participants have to rate how important they take these considerations to be. These statements are associated with different categories of moral reasoning. On the basis of the pattern of responses, researchers can assign participants a certain level of moral judgment.

[8] I rely here on Eisenberg's well-known account of prosocial moral reasoning (Eisenberg, 1986; Siegler et al., 2011). It is related to the original stages theory of Kohlberg, as well as the schema theory of the Neo-Kohlbergians (see especially Rest, 1999). My argument here does not depend on Eisenberg's particular account of moral development being correct. What it relies on is the idea, shared by all these accounts, that the young child's moral judgment is far from fully developed.

Besides being concerned with self-interest, many preschoolers also express concern about the physical, material and psychological needs of others, albeit in the most simple terms. This mode of reasoning is the predominant mode for elementary school children. Eisenberg associates it with a second level which she calls *needs-oriented*. Some elementary school children and many high school children come also to employ a more societally-oriented mode of reasoning. They come to reason in terms of stereotype images of good and bad, social norms, certain authorities, and the approval and acceptance of others. This *societally oriented reasoning* is associated with the third level of moral development.

In late childhood and adolescence, children begin to justify, in varying degrees, their moral judgments with references to perspective-taking and morally relevant emotions such as guilt (Siegler et al., 2011). For example, they may argue that a protagonist in a story should refrain from doing something to another character because she herself would also feel bad to be treated in that way had she been in the other's situation. Eisenberg calls this mode of reasoning *self-reflective empathic orientation*[9], and associates it with the fourth level of moral development. Empirical evidence suggests that for most persons this is the most mature stage of moral thinking that they reach, and that they employ it in combination with the modes of reasoning associated with earlier levels. It is presently thought that only a minority of people of high school age or older come to adopt a more abstract mode of reasoning that refers to internalized values, norms, duties, rights, or responsibilities (Siegler et al., 2011). Eisenberg associates this reasoning with a fifth and final level of moral development, which she calls the *strongly internalized stage*.

Findings on moral development thus reveal that the moral judgment of young children is qualitatively different from that of older children and adults. That means that even if the three-year-old's moral judgment would not require perspective-taking, which I have argued is not evident, it may be involved in more mature moral judgment. Interestingly, a central thesis of Kohlberg is that the development of moral judgment depends on advances in one's capacity for perspective-taking (Colby et al., 1983). This view is compatible with the recent developmental literature. As we saw above, while young children are able to understand the perspectives of others, this understanding is not fully matured. It is widely recognised that children tend to be self-centred. This changes in the course of development. Children

[9] It is called *self-reflective* because the child refers explicitly to how she would feel in the other's situation.

become both better at understanding perspectives different from their own and more prone to adopting them. Interestingly, the development of perspective-taking is not bound to childhood. It has been found that, just like the level of moral reasoning, the tendency to adopt perspectives different from one's own increases during adolescence and into adulthood (Eisenberg et al., 2005; Parker, Rubin, Erath, Wojslawowicz, & Buskirk, 2006). The development of perspective-taking and moral development thus appears to coincide. In the next section I shall present evidence that development of perspective-taking drives moral development.

It thus turns out that findings on the moral judgment of children are compatible with perspective-taking accounts of moral judgment. Nichols's second premise, that the moral judgment of 3-year-olds is in essence similar to that of mature moral agents, does not fit the empirical literature on moral development. Even if young children would not form moral judgments through perspective-taking, mature moral agents may do so. Whether they in fact do so is discussed further in §3.

What about findings on the moral judgment of individuals with autism? As was mentioned above, individuals with autism distinguish between conventional and moral norms, which is an important achievement in moral development. There are additional findings that support the conclusion that they have moral understanding despite their social cognitive limitations. As I mentioned before, Nichols takes this to show that perspective-taking, which is undoubtedly impaired in individuals with autism, does not have an important role in moral judgment.

It is questionable, however, whether individuals with autism attain a similar level of moral understanding as individuals without autism. Certain abnormalities in the moral judgments of individuals with autism have been observed. Although children with autism understand that actions stemming from ill motives are more culpable than actions from good motives, their understanding of this is less developed than that of children of the same verbal age (Grant, 2005). Adults with autism tend to judge accidental harm as less morally permissible and more blameworthy than normal subjects, and tend not to take intent into account when judging whether a person should be forgiven for moral transgressions (Moran et al., 2011; Roge & Mullet, 2011). Furthermore, whereas people generally tend not to praise an action if it was unintended, by far most individuals with autism tend to praise unintended actions with good consequences (Zalla & Leboyer, 2011). Related to these findings, it has been found that individuals with autism judge faux pas

scenarios—scenarios in which a speaker says something that might hurt or is unpleasant to the listener without intending this effect—differently from normal subjects. Individuals with autism are less able to detect the faux pas when confronted with such scenarios than normal subjects, appear often to be unaware of the false-belief and of the occurring emotional harm, and are unable to explain want went wrong (Shamay-Tsoory, Tomer, Yaniv, & Aharon-Peretz, 2002; Zalla, Sav, Stopin, Ahade, & Leboyer, 2009). They sometimes interpret speakers as if they intended the effect anyway and may then blame them accordingly (Zalla et al., 2009).

Why do these abnormalities occur? On a first view, it may seem that the reason that individuals with autism do not take the presence or absence of intent into account is simply because they have a hard time recognising it. Due to their social cognitive limitations, their ability to gather morally relevant information is impaired. However, that interpretation does not seem plausible with regard to these studies. Most of the above studies let participants form moral judgments about the behaviour of story characters whose intentions have been made explicit. Indeed, several of the above studies note explicitly that participants with autism acknowledged the presence or absence of intent in the scenarios. Furthermore, two of the studies report that individuals with autism performed just as well on verbal theory of mind tasks as normal subjects (Moran et al., 2011; Zalla & Leboyer, 2011). This indicates that even when individuals with autism notice the presence or absence of intent, it plays only a limited role in their moral thinking. As Zalla and Leboyer (2011) observe, "moral judgments of the agent's action in individuals with [high-functioning autism] appeared to be primarily affected by the intrinsic goodness/badness of the outcome and to a lesser extent by the evaluation of the agent's motives or psychological attitudes" (p. 14).

There is evidence for a more fundamental limitation in the moral understanding of individuals with autism. Whereas both children and adults readily provide justifications for their moral judgments, individuals with autism have great difficulty doing so (Grant, 2005; Shamay-Tsoory et al., 2002; Zalla et al., 2009). When asked to do so, they are more likely to reiterate provided information than to give reasons. In particular, in contrast to even very young children, they will not refer to harm that has been caused, nor to any relevant attitudes of the judged persons. On the basis of such findings, Frederique de Vignemont and Uta Frith, one of the leading experts on autism, have argued that it is doubtful whether individuals with autism have genuine moral understanding (De Vignemont & Frith, 2008). The moral judgment of

individuals with autism may derive from their general ability to learn normative rules. As Baron-Cohen, another expert on autism, writes: "the only information they are interested in is patterned, systemizable information" (p. 115). Individuals with autism are focused on patterns and skilled at recognizing them, much more so than ordinary people (Baron-Cohen, 2011). It is through identifying the patterns or rules of social interaction that they can develop some grasp of it. This ability to 'systemize', as Baron-Cohen calls it, may also underlie their grasp of moral rules (Baron-Cohen, 2011). Given their ability and urge for systemizing, it is not surprising that individuals with autism could come to understand that certain transgressions are less permissible, more serious, more culpable, and less authority-dependent than others.[10] However, that would not mean they understand the reasons that underlie these moral rules. As De Vignemont and Firth write:

> It is not surprising that individuals with ASD [Autism Spectrum Disorders] are sensitive to normative rules, given that these rules are the only way they have to cope with their lack of social intuitions. Still, it does not mean the rules they obey are nothing more than conventional for them. (De Vignemont & Frith, 2008, p. 280)

Taking the above considerations together, I conclude it is questionable that individuals with autism have a similar level of moral understanding as normal adults.

Let's take stock. I discussed two premises of Nichols's argument against the idea that perspective-taking has a significant role in moral judgment. I first argued that the premise that young children and individuals with autism cannot grasp perspectives different from their own is not warranted, particularly not with regard to children. It is thus not evident that their moral understanding does not depend on perspective-taking. I then argued that the premise that children and individuals with autism have a similar capacity for moral judgment as adults is unwarranted. Studies in moral development suggest there are important differences in moral understanding, which happen to coincide with differences in capacity for perspective-taking. That means that

[10] The experiment on which Nichols bases his claim that children with autism understand the difference between conventional and moral rules asked them to judge the permissibility of certain transgressions, the seriousness of the transgressions, and whether the transgression would be permissible if an authoritative figure would say it is so. In contrast to experiments on the ability of preschoolers to make this distinction (e.g. Smetana, 1985), individuals with autism were not asked to justify their judgments. While these other experiments found that even preschoolers justify their judgments in terms of harm that has been caused, the above-mentioned studies suggest that individuals with autism would not.

even if the first premise were correct, which I deny, perspective-taking may have a significant role in mature moral judgment.

2.3 Conclusions

This section discussed two challenges to the thesis that reasoning that involves perspective-taking has a significant role in moral judgment, the truth of which would enhance the Practicability Assumption. I have argued that empirical findings on moral judgment do not substantiate these challenges. Two conclusions are worth repeating. The first is that persons not only often draw moral judgments without first engaging in reasoning that involves perspective-taking, there also is reason to think that such reasoning commonly requires too much effort. The second is that this does not imply that such judgments do not depend in a significant way on reasoning that involves perspective-taking. Even when reasoning that involves perspective-taking does not have a synchronic role, it may yet have a diachronic role. This view is compatible both with findings on the role of intuitions in moral judgment, and with findings on the limited role of perspective-taking in the moral judgment of individuals with autism and that of children.

3 Support for a significant role for perspective-taking

Contrary to what certain moral psychologists have suggested, existing findings on moral judgment do not count against the idea that perspective-taking has a significant role in moral judgment. Although they suggest that reasoning that involves perspective-taking does not and cannot have a synchronic role in the formation of every single moral judgment, this conclusion is compatible with it often having a synchronic role and in other instances having a diachronic role instead. The present section shall consider whether there is evidence in favour of such reasoning indeed having these roles in moral judgment, which would in turn enhance the plausibility of the Practicability Assumption: it would support thinking that people have the capacity most crucially required for adopting the contract test as a moral guide.

I will thereto consider empirical findings concerning the relation between perspective-taking and moral judgment. First I will consider findings regarding the relation between competence in perspective-taking and moral judgment. After that I turn to findings regarding the relation between the exercise of the capacity for perspective-taking and moral judgment.

3.1 Moral judgment and perspective-taking competence

If people form moral judgments through perspective-taking, we may expect their competence in perspective-taking to be reflected in their judgment. Several studies suggest that competence in perspective-taking is indeed positively associated with moral understanding. I shall start by considering studies that look at moral judgment directly, then turn to studies that consider moral reasoning, and finish this subsection with studies that concern the relation between perspective-taking competence and moral behaviour.

First of all, differences in perspective-taking competence are reflected in moral judgments about intent and blameworthiness. For one, children who do not pass the false-belief test mistakenly judge that an agent who inadvertently hurts another person's feeling did so intentionally, and blame them accordingly (Killen, Mulvey, Richardson, Jampol, & Woodward, 2011). This is not surprising given that they may have been unaware of the agent's true motives. Perspective-taking may, however, also be required for taking intentions properly into account. As I mentioned in the previous section, even when aware of them, individuals with autism tend not to take motives of agents sufficiently into account when forming moral judgment about them. Several studies suggest this is because doing so also requires perspective-taking competence.

Ittyerah & Mahindra (1990) compared the perspective-taking skills and moral judgment of children at various ages. They measured both visual and cognitive perspective-taking. Visual perspective-taking was measured through the three mountain task, first introduced by Piaget and Inhelder (1967). In this task, participants are confronted with a toy man who is placed at various positions around three mountains. They are then asked to associate the position of the toy man with alternative pictures of the mountains. For cognitive perspective-taking, Ittyerah & Mahindra made use of a role-taking task first used by Flavell and colleagues (1968). Participants are asked to describe a sequence of cartoons, first from their own point of view, and then from the perspective of another person whom they know to have been shown only an abbreviated version of the same stimulus materials. Participants have thus more information than this other person, and an intrusion of this information in their descriptions of the other's perspective is evidence of a failure in perspective-taking. Finally, to measure moral understanding the researchers made use of a task in which participants were presented with stories in which a protagonist either lied to a person or injured an animal. For

example, they were told about Mary, who, when being asked by Peter to come and play, told him that she was sick, only to go shopping afterwards. The participant had to appraise the action on a seven-point scale, ranging from very very bad to very very good. After the participant had made an initial judgment, the investigator told the story again but added some additional information regarding the motives of the protagonist. For example, the participant was now told that Mary went shopping to buy a present for Peter's birthday, which was the next day.

Ittyerah & Mahindra (1990) found that among five- and six-years-olds, children who perform better on perspective-taking tasks put more weight on the motives of protagonists. Using different measures, Baird and Astington (2004) report similar results with regard to four and five year old children. They found that the extent to which children take intentions into account when judging whether an action is wrong and whether the actor should be punished was positively associated with the children's performance on a false-belief task. Such findings suggest that perspective-taking is not only involved in understanding intentions, but also in the reasoning processes that underlie moral judgment.

Developmental psychologists have studied the role of perspective-taking in moral reasoning also more directly, which constitutes the second line of research that I consider here. Selman (1971) investigated the relation between perspective-taking competence and the development of moral reasoning amongst 8- to 10-year-olds. To measure perspective-taking competence he made use of the previously mentioned cognitive perspective-task of Flavell and colleagues (1968), as well as an additional task in which the child had to predict another child's strategy in a game in which motives played an important role. To measure the development of moral reasoning he used Kohlberg's Moral Judgment Scale, an interview measure related to Kohlberg's theory of moral development. Selman found that children who performed better at the perspective-taking task were more likely to employ what Kohlberg calls conventional moral reasoning rather than the less advanced pre-conventional moral reasoning.[11] Note that this finding suggests that

[11] It is worth noting that subsequent analysis showed that the relation between perspective-taking and moral reasoning occurred only amongst children who have reached a medium mental age (who have average IQ with respect to their chronological age). Selman explains that this is because children of a lower mental age are not yet utilizing their newly acquired perspective-taking skills properly in moral judgement, whereas children of a high mental age had already developed this skill earlier.

development of moral judgment beyond the age of six also depends on perspective-taking competence.

In a follow-up that took place a year later with the children who did not display conventional moral thinking in the first study, Selman found that all of those who now displayed conventional moral thinking had improved perspective-taking competence as well, further supporting the idea that improvements in perspective-taking drive improvements in moral reasoning. Interestingly, a few of them had become more competent at perspective-taking but not at moral reasoning. Selman suggests these children had not yet sufficiently integrated their perspective-taking capacity in their reasoning.

That moral reasoning is affected by perspective-taking competence is also supported by studies on juvenile delinquents. Juvenile delinquents are significantly worse at perspective-taking tasks than their age-peers (Chandler, 1973; M. Lee & Prentice, 1988). Lee and Prentice (1988) found that they in addition display less advanced moral reasoning than their age-peers, as measured by items on Kohlberg's Moral Judgment Scale. Subsequent analysis revealed a relatively high correlation between performance on the perspective-taking task and moral reasoning, even when corrected for intelligence.[12] While there was an additional correlation between logical reasoning and moral maturity, this disappeared when the effect of perspective-taking was partialled out. In the words of the researchers, "role-taking appears to play a mediating role between logical cognition and moral reasoning" (Lee & Prentice, 1988, p. 135).

Not all studies on the relation between perspective-taking competence and moral reasoning report a connection. Eisenberg and Roth (1980) found no relation between the ability of 5-7 year olds to do Flavell's role-taking task and their level of moral reasoning. Inspired by the contradictory state of the existing evidence, Lane and colleagues (2010) performed a longitudinal study to gain more insight in the relation between perspective-taking competence and modes of moral reasoning. They measured the perspective-taking capacities of children in two waves: first when they were between 2 and a half and 4 years old, and again when they were 5 or 6 year old. They also measured their capacity to understand emotions, using a task that had children identify emotions of line-drawn faces and a task that had them identify the emotions of story characters in various situations. In the second wave of tests, the

[12] It should be noted that researchers typically correct for intelligence. I mention it here explicitly because the lower intelligence of delinquents may seem a plausible explanation of their less advanced moral reasoning.

researchers in addition measured children's capacity to coordinate their understandings of the mind and emotions. For this purpose, children were told about a protagonist who tried to hide his emotions, and were asked to identify both what he tried to look like and how he actually felt, again using line-drawn faces. In this second wave their level of moral understanding was also assessed, using similar tests as Eisenberg and Roth (Eisenberg & Roth, 1980). Besides having to tell what the protagonist in a moral scenario should do and why, children were also asked to give reasons for why the protagonist may do the opposite. Reasoning responses were coded in terms of the modes of reasoning distinguished in Eisenberg's framework of moral development (see §2.2).

Lane and colleagues (2010) found that perspective-taking capacities did predict the level of moral reasoning, even when adjusting for intelligence. Children between 2.5 and 4 years old who expressed a better understanding of false-beliefs were more likely to take into account the psychological needs of others, and children who expressed a better understanding of emotions were more likely to reason in terms of the physical and material needs of others. In contrast to Eisenberg and Roth (1980), Lane and colleagues found that children who performed better at both tests were less likely to rely on hedonistic reasoning, the most simple mode of moral reasoning. In addition, they found that children who performed well on coordination task in wave 2, arguably the most difficult perspective-taking task, were more likely to engage in socially-oriented reasoning, the most advanced mode of reasoning tested.[13]

The last noteworthy line of research concerns the relation between perspective-taking and moral behaviour. Several studies have reported there to be a weak to modest positive relation between moral judgment and moral behaviour (Eisenberg, 1986; Epley & Caruso, 2004; Underwood & Moore, 1982). If people form moral judgments through perspective-taking, we may thus also expect there to be a positive relation between perspective-taking competence and moral behaviour. Put differently, were we to find that

[13] Why did Lane and colleagues (2010) find a positive relation between perspective-taking and level of moral reasoning where Eisenberg & Roth (1980) did not? Lane et al (2010) suggest this may be because they relied on different measures. First, Lane and colleagues measured perspective-taking differently, attending also to affective perspective-taking. Given that moral dilemmas often involve emotional content, such as persons being hurt, affective perspective-taking may be involved in more advanced moral reasoning. Second, participants in Lane and colleagues' study had their reasoning measured in two different ways. In contrast to the study of Eisenberg and Roth, participants had not only to provide reasons in support of their decision but also reasons against it. When Lane and colleagues took only the former into consideration, they found essentially no relationship, like Eisenberg and Roth.

perspective-taking competence increases moral behaviour, this would provide further support for thinking that perspective-taking affects moral judgment.

There is evidence for this. Underwood and Moore (1982) conclude on the basis of a meta-study encompassing 10 studies that there is a reliable positive correlation between perspective-taking competence and helpfulness. Among these studies is the finding that 6-year-olds who receive perspective-taking training become more inclined to share with other children. There is also evidence that perspective-taking competence plays a role in fairness-related behaviour. Takagishi and colleagues (2010) measured the relation between false-belief understanding and behaviour in Ultimatum games amongst 5-6 year olds. One child was told to propose a distribution of 10 candies between herself and another child, with whom she had physical contact, whereas the other was told she could either reject or accept the offer. They found that 19 out of 23 children who passed the false-belief task chose to share the candies evenly, whereas only 4 out of 11 children who did not pass the task offered such a fair distribution.

Besides promoting prosocial behaviour, perspective-taking competence may also reduce antisocial behaviour. Chandler (1973) found that increases in perspective-taking competence may reduce antisocial behaviour. After giving delinquent children a 10-week, half a day a week, perspective-taking training, they performed much better on perspective-taking tasks. Chandler also found that they subsequently committed fewer criminal offences: police reports revealed that individuals who received this training committed on average approximately half as many (known) delinquencies during a period of one-and-a-half years after the treatment than delinquents who had been assigned to either a control or a placebo group. Assuming that the effects of perspective-taking competence on behaviour are mediated by judgment, this finding again supports that perspective-taking is involved in moral judgment.

It may be responded at this point that there is another viable explanation for why perspective-taking competence would improve behaviour. Rather than affecting a person's judgment, perspective-taking may only affect a person's *motivation*. While there is no knock-down argument against this explanation, there are reasons to prefer the above explanation. First, this alternative explanation would have the counterintuitive implication that children in the above experiments would, despite acting very differently, agree in their judgments regarding what they have reason to do. Second, there are other different lines of research, we saw above, that suggest perspective-taking competence does affect moral judgment. The above explanation is supported

by these findings. I discuss the competing explanation more extensively in the next subsection.

The above three lines of research support the view that perspective-taking has a significant role in moral judgment. More precisely, the positive association between perspective-taking competence and moral reasoning suggests that perspective-taking is involved in reasoning processes that underlie moral judgment. Less evident, however, is whether this role is synchronic or diachronic. Do people form moral judgments through occurrent reasoning that involves perspective-taking, or is it involved in the development of a moral understanding that is reflected in judgment?

There is reason to think it is involved in both ways. First, there is an argument to be made that if reasoning involving perspective-taking is to have a diachronic role in moral judgment, it must also have had a synchronic role. The idea is that if moral judgments depend on a moral understanding that has been developed or shaped through reasoning that involves perspective-taking, there must have been occasions on which this shaping took place, and in which such reasoning thus had a synchronic role. Take for example the child who has has come to realise that the extent to which a person should be punished for a transgression depends on the person's motives. The argument would be that if her moral insight depends on improved reasoning that involves perspective-taking, as Ittyerah & Mahindra (1990) maintain, the child must have at some point arrived at this conclusion through such reasoning. This is one argument for thinking that perspective-taking has a synchronic role in moral judgment; in the following section I shall provide another.

Second, there is an argument to be made that if reasoning that involves perspective-taking has a synchronic role in moral judgment, it is likely to also have a diachronic role. Say, again, that a person through reasoning that involves perspective-taking at a certain point in time concludes that a person who unintentionally hurts a third party should not be blamed to the same degree as a person who intentionally hurts a third party. In this case, such reasoning has a synchronic role. Say now that at a later point in time this person is confronted with a situation in which another hurts a third party unintentionally. It seems plausible that, due to the conclusion arrived at in the past, the person can judge that the other should not be judged too harshly without having to engage in reasoning that involves perspective-taking again (although perspective-taking may be needed to recognise it was done unintentionally). If so, the reasoning which has taken place earlier would now have a diachronic role.

We thus have one argument stating that if reasoning that involves perspective-taking has a diachronic role it must have a synchronic role, and another stating that if such reasoning has a synchronic role it is likely to also have a diachronic role. Given that the findings indicate that reasoning involving perspective-taking has at least one of these roles, I conclude it is likely to have both of them. Note, however, that the findings are consistent with it having one of these roles much more frequently than the other. Indeed, given the effort that such reasoning involves, it is not unreasonable to think that most moral judgments depend on it only diachronically.

Let me finish by observing that this conclusion on the role of reasoning involving perspective-taking in moral judgment supports the previous section's proposal (§2.1). I mentioned there that moral contract theorists such as Gauthier and Scanlon may propose that persons use the contract test to internalise moral principles on which they can rely when their cognitive resources are low. If the moral judgments of a person are already diachronically based on reasoning that involves perspective-taking, the above findings provide support for thinking that persons can indeed through reasoning that involves perspective-taking shape their future judgments.

3.2 Moral judgment and the exercise of perspective-taking

Persons will now and then be confronted with new or difficult moral cases. For the contract test to be a proper moral guide, persons must be able to use it to evaluate such cases. It must thereto be able to have a synchronic role. Although the previous subsection concluded that in order to develop moral understanding persons must at some point have engaged in reasoning that involves perspective-taking, this is consistent with persons not usually being able to employ such reasoning. In order to have reason for thinking that persons can apply the contract test to evaluate new or difficult cases, I will examine whether there is evidence that, apart from moral development, reasoning involving perspective-taking has a *regular* synchronic role in moral judgment. If so, we would have reason to think persons can also rely on the contract test in the relevant situations.

If perspective-taking that involves reasoning has a regular synchronic role in moral judgment, we may expect to find that the judgments of persons are affected not only by their competence in perspective-taking but also by whether they exercise this capacity or not. Related to this, we may expect there to be a relation between a person's tendency to engage in perspective-

taking and what moral judgments she tends to form. I start with some findings regarding the second relation.

A series of studies have found that persons who, according to self-reports, are more likely to consider the perspectives of others tend to display a higher level of moral reasoning when asked to justify their moral judgments (Eisenberg et al., 2005; 2002; Eisenberg, Zhou, & Koller, 2001; Myyrya, Juujärvi, & Pesso, 2010). In order to measure the willingness to adopt the perspectives of others, researchers typically make use of Davis's Interpersonal Reactivity Index (IRI), a questionnaire meant to measure affective and cognitive dispositions central to empathy (Davis, 1983). It requires persons to rate themselves on items such as "I sometimes try to understand my friends better by imagining how things look from their perspective" and "When I'm upset with someone, I usually try to put myself in their shoes for a while". In a study of adolescents, Eisenberg and colleagues (2001) found a correlation of 0.36 between scores on the IRI and scores on the Prosocial Reasoning Objective Measure, a pencil and paper measure to assess a persons level of moral reasoning. Breaking the results up for various modes of moral reasoning, it was found that less advanced modes of moral reasoning, such as hedonistic reasoning, are negatively associated with the tendency for perspective-taking, whereas the most advanced mode of moral reasoning is positively associated with this tendency. Put differently, people who are inclined to adopt the perspectives of others were found to have a higher level of moral reasoning. Myyrya and colleagues (2010) found a similar pattern with respect to the Defining Issues Test, which measures participants' level on Kohlberg's scale of moral development: amongst university students, a tendency for perspective-taking was found to be negatively associated with the pre-conventional and the conventional level of moral reasoning, but moderately positively associated with the more advanced post-conventional level.

Why would persons who are more inclined to engage in perspective-taking have a higher level of moral reasoning? One plausible explanation is that higher levels of moral reasoning actively involve perspective-taking. If this were the correct explanation, the above findings support that reasoning involving perspective-taking has a regular synchronic role in moral judgment. The findings are, again, also consistent with it being involved only diachronically, however: persons who are more inclined to engage in perspective-taking may have, due to this tendency, developed a higher level of

moral reasoning, without perspective-taking being actively involved when they reason.

To find out whether perspective-taking has a regular synchronic role in moral judgment or not, we need studies that measure the moral judgment of persons we know to differ in their exercise of perspective-taking when forming these judgments. Surprisingly, such studies have not, as far as I know, been undertaken. Studies have shown that when people consider the perspective of a person belonging to a different group, they become less prone to apply stigmas or stereotypes (Batson et al., 1997b; Galinsky & Moskowitz, 2000). Although this is evidence for a synchronic role for perspective-taking, and as such relevant for our purposes, these judgments are not the sort of moral judgments we are primarily interested in.

What we do have are studies regarding the relation between the exercise of perspective-taking and moral *behaviour*. Such studies do not provide direct evidence of the effect of perspective-taking on moral judgment. However, in so far as it may be assumed that the effects of perspective-taking are mediated by moral judgment, such studies may provide indirect evidence about the effects of perspective-taking on moral judgment. After describing some of these findings, I will briefly discuss the plausibility of this 'mediating assumption'.

Several studies report that engaging in perspective-taking tends to increase prosocial behaviour. To start with, persons have been found to become more inclined to help others after adopting their perspectives (Batson & Shaw, 1991; Maner et al., 2002; Oswald, 1996). For one, Oswald (1996) found that, compared with participants instructed to focus on technical aspects of a video displaying a person in need, participants who are instructed to discern the thoughts and feelings of the videotaped person report to be more willing to help persons in a similar situation.

There is also support for a relationship between the tendency for perspective-taking and prosocial behaviour. Although an initial study found no relation between perspective-taking as measured by the IRI and helping (Davis, 1983), later studies have (e.g. Eisenberg et al., 1989). In addition, persons who are more inclined to perspective-taking are more likely to be engaged in long-term voluntary work (Unger & Thumuluri, 1997).

Besides being positively associated with prosocial behaviour, perspective-taking is also negatively related with antisocial behaviour. Persons who are more inclined to perspective-taking respond less aggressively in experiments constructed to arouse aggression (Richardson, 1998). Furthermore, the willingness to consider other points of view is associated with less destructive

behaviour in close relationships (Arriaga & Rusbult, 1998). When conflicts arise in a close relationship, several kinds of responses are possible. A robust finding is that when, inevitably at certain points in a relationship, one individual does engage in destructive behaviours—yelling, threatening to leave, avoiding discussion of critical issues, reducing interdependence—"couple functioning is enhanced to the extent that the partner (a) inhibits the impulse to respond in kind [...] and (b) instead reacts in a constructive manner" (Arriaga & Rusbult, 1998, p. 928). Arriaga and Rusbult found that people who are disposed to adopt one's partner's perspective are less likely to act destructively in the relationship, and more like to respond constructively. In a follow-up experiment, they also found that participants who were confronted with an imaginary scenario in which their partner acted destructively showed significantly more constructive and less destructive responses after they had been manipulated into looking at the situation from their partner's perspective. Perspective-taking thus affected their judgment synchronically.

The negative relation between the tendency for perspective-taking and antisocial behaviour is also confirmed by studies on offenders. A recent meta-analysis reports a relatively strong relationship between offending and lack of empathy, including a willingness for perspective-taking (Jolliffe & Farrington, 2003).[14] While the researchers do not connect this explicitly to moral judgment, they do conclude that criminal behaviour is at least in part explained by limited perspective-taking. Furthermore, other studies have reported that delinquents, at least during adolescence, have less mature moral judgment than their age-peers (Nelson, Smith, & Dodd, 1990; Stams et al., 2006).[15]

Analogously, psychopathy appears to be associated with a reduced willingness for perspective-taking. Lack of empathy with others is a key characteristic of psychopaths (Hare, 1993). Studies have found that undergraduate students who score high on a psychopathy measure also report to be less willing to engage in perspective-taking (McIlwain et al., 2012;

[14] Not all individual studies find this. A possible explanation lies in the fact that the IRI and other empathy tests rely typically on self-reporting. It is not uncommon for individuals low in empathy to think of themselves as being quite empathic (Baron-Cohen, 2011)

[15] Barriga and colleagues (2009) report that amongst juvenile delinquents competence in moral judgment was associated with a greater willingness to adopt the perspective of others. An interesting additional finding of theirs is that delinquent youth with little empathy tend to engage in self-serving cognitive distortions. As the researchers point out, this fits with the idea that self-serving cognitive distortions can be used to neutralise empathy.

Mullins-Nelson, Salekin, & Leistico, 2006). Again there is reason to think that the reduced tendency for perspective-taking resonates in moral judgment. Analysing participants' verbal responses on a moral dilemma (i.e. the Trolley Problem), McIlwain and colleagues found that participants who scored high on psychopathy did not refer to the perspectives of persons involved, in contrast to normal subjects.

Taken together, the above studies show convincingly that the exercise of perspective-taking affects social behaviour. On the assumption that perspective-taking has these effects by being involved in reasoning processes that affect behaviour through affecting moral judgment first, such studies provide evidence for perspective-taking having a regular synchronic role in moral judgment. In that case, they also provide evidence for thinking persons can apply the contract test to evaluate difficult and new cases that they face. More precisely, given that participants in the above experiments are actual persons under conditions of medium rather than high cognitive resources, they provide evidence persons can apply the contract test under circumstances of everyday life in which their cognitive resources are not too low.

I will finish this subsection by discussing the plausibility of the mediating assumption on which this conclusion depends. It is not supported by the studies themselves: they provide too little insight into how participants arrive at their choices. My approach will therefore be different. First I will argue that the explanation that perspective-taking affects behaviour through judgment is a better explanation than what I take to be the most viable alternative. After that I will briefly consider whether perspective-taking may synchronically affect moral judgment in some other way than through being involved in reasoning. Although I will not be able to resolve this final issue, I shall argue this does not prevent us from drawing the conclusion that reasoning involving perspective-taking *can* have a regular synchronic role in moral judgment.

The alternative explanation on which I concentrate, already introduced in the previous subsection, is that perspective-taking affects behaviour not through moral judgment but through motivation alone. I mentioned there two considerations that count against this explanation. First, it has the counterintuitive implication that participants in the above studies choose differently but agree in their judgment about what they have reason to do. The second consideration I introduced was that other lines of research do show that perspective-taking affects moral judgment. However, given that the these lines of research are consistent with perspective-taking having only a

diachronic role in moral thinking, this consideration does not count against this alternative explanation of the synchronic effect.

There is an argument in favour of the alternative explanation. It is a common view in moral psychology that emotions and moral motivation are associated. There is evidence that adopting a person's perspective tends to generate empathic and sympathetic feelings for that person (Batson et al., 1997b; Pizarro, 2000; Pizarro & Bloom, 2003). Indeed, in research on empathy, perspective-taking manipulations are often used to get participants to have empathic concern for others. It is plausible that such prosocial emotions play a role in moral behaviour. For example, Oswald (1996) found that perspective-taking was most likely to lead to helping when it resulted in affective empathy. More dramatically, Coke, Batson & McDavis (1978) found that when people who engage in perspective-taking are led to believe, falsely, that the physiological arousal they experience due to perspective-taking is the result of an arousal-inducing pill, they are much less likely to decide to help. The researchers take this to show that perspective-taking without an affective response does not increase helping behaviour.

While these findings are relevant for choosing between the two explanations, they actually count in favour of the original explanation. Not only is perspective-taking causing prosocial emotions towards others perfectly compatible with it affecting moral judgment, there is emerging evidence of a close relation between moral judgment and such emotions (Eisenberg et al., 2001; Haidt, 2001; Nichols, 2004; Pizarro, 2000). The psychologist David Pizarro (2000) describes several ways through which emotions such as empathy and sympathy may affect moral judgment, of which I will briefly mention three. First, emotions may have a signalling function, triggering a person to form a moral judgment. As Pizarro says, "when empathy arises in the presence of a distressed other, the empathic response cues the individual to the possibility that a morally relevant event is taking place" (p. 360). Second, emotions may have a correcting function, leading a person to reconsider a previously formed moral judgment. Third, as we saw in §2.1, emotions may sometimes directly lead to a moral judgment. Besides triggering a person to form a moral judgment or correct an existing moral judgment, a feeling of sympathy may lead him to instantly conclude that he ought to help.

Daniel Batson and colleagues have done some interesting studies regarding the interplay between emotions and moral judgment. In a series of experiments, Batson and colleagues found that when the value that a person assigns to another's well-being increases, so do his empathic feelings towards

that person (Batson, Turk, Shaw, & Klein, 1995). What may be more surprising, when participants were manipulated into feeling more empathy for another (through perspective-taking instructions), so did the value that they assigned to the other's well-being. Interestingly, this latter effect remained even after affective empathy again decreased. Batson and colleagues take these findings to show that people take their empathic responses to others as indications of the degree to which they care about others. Put differently, prosocial emotions affect judgment.

The previously mentioned study of Coke, Batson & McDavis (1978) provides reason for thinking that without this effect of emotions on judgment, emotions may not even influence a person's behaviour. As I mentioned before, participants in this study were manipulated into experiencing sympathy for another person in need through being instructed to adopt that person's perspective. Before this, however, some participants were given a pill that they were told, falsely, would lead to physiological arousal. After the perspective-taking task, participants were asked whether they were willing to help the person in need. The researchers found that persons who had not taken the pill were quite willing to help. Persons who did take the pill, on the other hand, were much less willing to help. Apparently, even though they experienced similar physiological arousal to the other's situation as those who did not take a pill, these participants were not inclined to help because they attributed the arousal they experienced to having taken a pill. They judged the situation differently, which in turn affected their motivation.

All this does not sit well with the alternative explanation. This explanation holds that perspective-taking generates prosocial emotions that only affect motivation and not judgment. However, the above suggests that the relation between judgment and motivation is much more intimate. As the original explanation *is* compatible with this intimate relation, it is the better explanation. I therefore conclude that the mediating assumption that perspective-taking affects behaviour through affecting moral judgment is plausible.

There is one remaining issue. It may thought that the above findings do not show that perspective-taking affects moral judgment because it is involved in *reasoning*. They are just as well consistent, it may be argued, with the view that perspective-taking generates emotions or intuitions that subsequently lead to moral judgments, as some of the moral psychologists introduced in the previous section would suggest, without reasoning coming into the picture.

This issue I cannot resolve here. The above studies provide insufficient information about how exactly participants arrived at their choices. Fortunately, the issue does not need to be resolved for answering the main question. I investigated whether persons *do* regularly form moral judgments through reasoning that involves perspective-taking first and foremost in order to find out whether they *can* do so, which is what the Practicability Assumption requires. I take it as a given that persons do regularly form moral judgments through reasoning. This section showed that perspective-taking as well regularly has a synchronic role in moral judgment. If we add to this that perspective-taking is at least sometimes involved in reasoning, as shown by several studies presented in this section, it is highly plausible that reasoning *involving* perspective-taking *can* have a regular synchronic role.

4 Conclusions

The Practicability Assumption states that persons can come to rely on the contract test as a moral guide. As the contract test is a reasoning procedure that involves considering perspectives different from one's own, this assumption requires that persons can, under appropriate circumstances, form judgments through such reasoning. I examined this requirement on the basis of findings regarding how persons actually form moral judgments. More precisely, I examined to what extent empirical findings support the view that we already form moral judgments through reasoning that involves perspective-taking. This view being correct would make it plausible that persons can adopt the contract test as a moral guide.

Several conclusions can be drawn. A first conclusion concerns the circumstances under which persons can apply the contract test. It is unlikely that persons can do so when their cognitive resources are low. Both reasoning and perspective-taking are effortful. However, there is reason to think that persons are able to apply it when their cognitive resources are medium or high. Empirical findings reveal that a capacity for perspective-taking is part of the machinery that underlies moral judgment. More precisely, there is evidence for both a synchronic and a diachronic role for reasoning that involves perspective-taking: that persons sometimes form moral judgments through reasoning that involves perspective-taking, and that the development of moral understanding depends on such reasoning. It is therefore plausible, I concluded in the previous section, that persons are regularly able to form moral judgments through reasoning that involves perspective-taking.

We may also draw a second conclusion about what sort of moral guide the contract test may be. Given that the contract test cannot be applied when cognitive resources are low, it cannot function as a decision procedure that persons consciously follow for all their moral choices. But neither do moral contract theorists such as Gauthier and Scanlon so intend it. As we saw, they intend it as an instrument to assess the moral principles on which we may act. I have argued that we have reason to think that persons can apply it under many of the situations of everyday life.

That persons cannot apply the contract test when their cognitive resources are low poses a potential problem. How can persons evaluate possible courses of action if the contract test is not available as a guide? A possible solution would be that persons prepare themselves for this sort of situation by internalising moral principles. That reasoning involving perspective-taking appears to have a diachronic role provides support for the idea that persons can indeed use the contract test in this way. By applying the contract test under favourable conditions, drawing conclusions about the justification of practices and types of actions, people may prepare themselves for situations in which they have insufficient cognitive resources to apply it. I will have more to say about this in Chapter 5.

It is important to emphasise that this chapter has only been concerned with the question of whether and under what conditions we can come to base our judgments on the contract test, not whether we can come to apply it *adequately*. This is what the next chapter will discuss.

4

Perspective-Taking Accuracy and the Contract Test

> I was in the House of Representatives for 25-and-a-half years, and I disagreed with the occupant of the White House, whether he was a Democrat of a Republican. I used to say: 'How can he be so autocratic, so dictatorial, why does he not understand that the congress is doing the right thing?' Well, when I moved from one end of Pennsylvania Avenue to the other end and occupied the Oval Office, my perspective changed significantly. And then I would look down at the congress and say 'What are those people doing over there? How can they be so irresponsible?' (Gerald Ford, 38th President of the United States, 2002)

1 Introduction

Moral contract theorists such as Gauthier and Scanlon assume, I have argued, that persons can rely on the contract test as a moral guide. In support of this Practicability Assumption, the previous chapter concluded that persons can under appropriate circumstances form moral judgments by applying the contract test. This, however, is just one of two requirements that must be met if the Practicability Assumption is to be satisfied. As I explained in Chapter 2 (§5), it must also be the case that persons can use the test *adequately*: that it is determinate and correct-usable to an appropriate degree. The present chapter investigates the plausibility of this aspect of the Practicability Assumption in the light of findings on perspective-taking.

As I explained in Chapter 2 (§4), whether we can use the contract test adequately depends to an important extent on how skilled we are at perspective-taking. The contract test can be characterised as a multi-step procedure, and at least two of its steps include perspective-taking.[1] The first of these steps is to identify the relevant perspectives. These are the points of view of persons that would in some way be affected by the general acceptance or

[1] In this chapter, as in Chapter 2 (§3), my characterisation of the contract test is closest to Scanlon's test. But as all contract tests require a correct understanding of alternative perspectives (2§4), most of what I say also concerns other contract tests, such as Gauthier's.

rejection of the principle under consideration. The second step is to understand exactly what implications acceptance or rejection of the principle has for persons in these situations. Or, to borrow a phrase from Scanlon, it is to identify 'objections' that persons occupying these points of view may voice against acceptance or rejection of the principle in question. Clearly, going through these two steps, whether it is implicitly or explicitly, requires more than just a capacity for perspective-taking that can be drawn upon for moral judgment. It is one thing to be able to consider other points of view; it is quite another to be able to identify the appropriate points of view and judge correctly how they would be affected by a principle. Applying the contract test adequately requires a certain *accuracy* in perspective-taking.[2]

Can people achieve sufficient accuracy in perspective-taking to become able users of the contract test in an everyday setting? As in the previous chapter, I will attempt to answer this question by considering studies on our capacity for perspective-taking. Given that only a few studies have considered our accuracy for perspective-taking in the course of forming a moral judgment, I shall mainly rely on studies regarding perspective-taking in a non-moral context. Although not ideal, this does not need to pose a problem. As I described before (2§4), there are important similarities between the kind of perspective-taking required by the contract test and other kinds of perspective-taking. Indeed, it is plausible that persons have a general capacity for perspective-taking that they can use for various purposes, including applying the contract test. In so far as psychological studies provide information about this general capacity, they also provide information relevant for evaluating our ability to apply the contract test adequately.

Empirical findings may affect the plausibility of this aspect of the Practicability Assumption in various ways (2§5). If studies show persons to be accurate perspective-takers, they provide support for the assumption that we can apply the contract test adequately. On the other hand, were they to indicate that persons tend not to be sufficiently accurate to apply the test adequately, they reveal a lack of support. Moreover, studies may provide findings that count *against* the plausibility of the Practicability Assumption. They may give us reason to think that the reported inaccuracy lies in certain limitations intrinsic to our ability for perspective-taking.

---

[2] As I mentioned in Chapter 2 (the final paragraph of §4), besides perspective-taking there are other capacities required for applying this test adequately. For one, it requires the capacity to judge whether implications count as reasons.

It is more than just a logical possibility that findings on our ability for perspective-taking would count against the Practicability Assumption. Social psychologists have argued that empirical findings reveal us to be inaccurate perspective-takers. As Nicholas Epley (2008) writes in a recent overview of the empirical findings, "people are fairly impressive mind readers in some instances and undeniably terrible in others" (p. 1456). Studies show that persons tend to project their own mental attributes onto others and as a consequence generate egocentrically biased interpretations of the points of view of others. Epley and his colleague Eugene Caruso (2004) argue that egocentricity is also reflected in the quality of our moral judgment. Research on perspective-taking shows that we often fail to consider events from perspectives different from our own, and that when we do consider them, we tend to overestimate the agreement with our own point of view. Epley and Caruso claim that this egocentric bias makes us "egocentric moral reasoners". Moreover, they suggest that there is little we can do to change this. As they write, "once a person is given a particular perspective on the world, it appears inevitable that this perspective will influence one's judgments, behavior, and moral reasoning" (p. 182).

This view fuels a concern for the Practicability Assumption. It suggests we may not be good enough at applying the contract test. Not only does it suggests we are prone to make mistakes when applying it, but also that there may be little room for improvement. As I explained before (1§3), this may have implications for the plausibility of contract theory as a moral theory. If persons have reason to think they cannot apply the contract test adequately, they have no reason to trust the conclusions they draw using it. As such, they cannot rely on it as a moral guide. Furthermore, they may also not have reason to adopt the particular moral principles defended by contract theorists: even if they make a convincing case that a principle would be the object of hypothetical agreement, we would have reason to distrust our judgment that they do.

This concern will be addressed in the present as well as the following chapter. In this chapter I will concentrate on studies that have been taken to show that we tend to be inaccurate perspective-takers. A crucial question with regard to these is whether they reveal our capacity for perspective-taking to be limited in certain ways. The following chapter will build on the findings of this chapter to discuss whether and in what ways we can account for such limitations. Put differently, it will be concerned with the question of whether and how we can become better at applying the contract test.

I should note that this chapter concentrates on the question of whether persons with the kind of perspective-taking accuracy that actual persons typically have can apply the contract test adequately by themselves, without communicating or collaborating with others. The next chapter will consider whether persons can improve their perspective-accuracy by gathering additional information or applying the test in collaboration. But in order to know whether persons need to adopt such methods it should first be considered how able we are at applying the test without them.

My discussion of the empirical studies is ordered on the basis of the two steps of the contract test distinguished earlier. The following section concentrates on what studies on perspective-taking accuracy show about our ability to identify points of view that must be taken into account when applying the contract test. The third section focuses on what such studies reveal about our ability to understand what objections may be posed from these points of view. It may be worth noting that, as there is no clear-cut distinction between these two steps of the contract test, the discussions of the second and the third section will not be wholly independent.[3] In the fourth and the fifth section I discuss what the findings of the previous sections mean for our ability to apply the contract test adequately.

2 Identifying alternative standpoints

Finding out whether a given principle for the general regulation of behaviour is one that everyone has reason to agree to requires first of all identifying alternative points of view with regard to the principle other than one's own present perspective. These may be the perspectives of actual others who would in some way be affected by the principle, or certain representative standpoints. I will in this section consider what empirical findings on perspective-taking show about our ability to identify alternative viewpoints.

[3] While there is a difference between identifying an alternative point of view and understanding it, they are closely related and in practice may often coincide. Clearly, understanding another point of view implies having identified it as an alternative point of view. But it is also the case that one cannot identify another point of view without having any understanding of it. In order to identify a different perspective regarding a given principle, I must be aware of differences between myself and occupants of that other point of view. Identifying another point of view thus presupposes some grasp of it. Given this interdependence, certain findings on perspective-taking may be relevant in both discussions. As my main interest is our ability to engage in the kind of perspective-taking involved in the contract test rather than these separate steps of the contract test, this overlap does not pose a problem.

Whether a person has identified another perspective is not easily observed directly. But it can be observed indirectly. Take the false-belief task discussed earlier. That a child predicts correctly that Maxi searches for his chocolate in the box where he left it rather than in the cupboard reveals that the child has recognised Maxi's different perspective. Similarly, if a person's judgment about how others perceive her is different from how she perceives herself, this reveals a sensitivity to other viewpoints. However, if a child does not account for Maxi's false belief when predicting his behaviour, or if a person's judgment about how others perceive her is affected by information only available to herself, this does not show a sensitivity to other perspectives. In contrast, one plausible explanation of such findings would be that alternative viewpoints have not been identified. Findings on our tendency to take into account other perspectives when forming judgments can thus reveal something about our ability to identify alternative points of view.

I start by presenting some negative findings, which suggest that we often do not take into account other viewpoints. In order to find out what these show about our ability to identify other points of view I consider how they should be explained. I shall relate this to our competence in applying the contract test, but leave most of the discussion for sections 4 and 5.

2.1 We often do not take other viewpoints sufficiently into account

Several lines of research in social psychology show we often do not take into account the particular perspectives of others when forming judgments about them, even though they are relevant for our judgment. First, they show we are prone to rely on stereotypes instead. When making judgments about an individual, people often draw on general propositions about groups of people that share a certain characteristic with that individual rather than to draw on that person's particular point of view (Galinsky & Moskowitz, 2000). Even if there is typically some truth in such stereotypes, and that as such they can teach us something about another's perspective, their accuracy is limited and when applied to individuals they can lead us to ignore many properties of individuals that do not fit the stereotype.

Second, a growing literature reveals that when making judgments about how others perceive us, we often presume them to have information that is in fact only available to ourselves. As Epley (2008; 2004) describes in overviewing the literature, people have been found to overestimate the extent to which others notice and attend to their behaviour (Gilovich, Medvec, & Savitsky, 2000; Savitsky, Epley, & Gilovich, 2001), the extent to which their

internal states are transparent to others (Gilovich, Savitsky, & Medvec, 1998), the extent to which others can recognise and hold nuanced impressions about their personality traits (Vorauer & Ross, 1999), and the use that others will make of information about their past when forming impressions about them (Chambers, Epley, Savitsky, & Windschitl, 2008). They also overestimate the extent to which others identify variability in their performance over time (Epley, Savitsky, & Gilovich, 2002), and the extent to which others will give them credit for specific tasks performed within a group (Kruger & Gilovich, 1999; Ross & Sicoly, 1979). Epley describes these findings as revealing an egocentric bias in our social judgments that is the result of information about ourselves—about our behaviour, our thoughts, and our contributions—being more readily available than information about others.

Third, there is evidence that our interpretations of the utterances of others are often affected by privileged information. Keysar and colleagues (2000) investigated the eye movements of people who were following instructions to manipulate objects given by another person. While this speaker could see some of the objects that participants could see, other objects were hidden from his sight. Interestingly, eye fixation data demonstrated that participants did not restrict the search for those objects that the speaker could see. More dramatically, follow-up studies found that even though participants were aware that a speaker did not know the identity of an object that they were carrying in a bag, they often interpreted descriptions of a similar object that was visible to both of them to refer to this hidden object; sometimes going so far as to comply with the instruction by moving the object in the bag rather than the mutually visible object (Epley, 2004; Keysar, Lin, & Barr, 2003). This occurred even when participants believed the director had a false belief about the object in the bag. In another series of studies, Keysar and Henly (2002) showed that people draw not just on privileged information when interpreting the utterances of others, but also when they consider the meaningfulness of what they say themselves. Speakers were found to underestimate the ambiguity of what they say and to overestimate how effective they were in conveying an intended message. Apparently, we implicitly assume others to have access to the same privileged information we have access to (Kawada, Oettingen, Gollwitzer, & Bargh, 2004).

These lines of research indicate that we are prone to fail to take into account alternative points of view relevant for our judgment. While the above studies do not concern moral issues, there is no reason to believe the findings

would not generalise to points of view relevant for our moral judgments.[4] Studies that find an egocentric bias in our judgments of fairness provide further support for this generalisation. Messick and Sentis (1979) found that when participants are asked whether it would be fair to give equal pay to two persons if one of them has worked for 10 hours and the other for 7, they are more likely to state it to be fair if they are themselves assigned the role of the person who worked for a shorter period. A similar role effect was found among participants who did not opt for equal payment. If they themselves were given the hypothetical role of the person who worked 10 hours rather than that of a person who received $25 dollars for 7 hours of work, participants took a little over $37 dollars to be a fair amount for themselves. However, when roles were reversed, they stated that the other deserved only $32.79.

That our judgments of fairness can be biased by our present position has been demonstrated even more dramatically by a mock trial study by Loewenstein and colleagues (1993). Participants who were randomly assigned the role of plaintiff or defendant in a hypothetical court case turned out to have very different perceptions of what a fair settlement would be. While plaintiffs on average considered 37 thousand dollars a fair settlement, defendants believed it should be 19 thousand. Interestingly, their roles even affected their estimations of what the judge would decide. Whereas plaintiffs thought the judge would rule 39 thousand, defendants believed it would be 24 thousand. Apparently, their predictions did not sufficiently take into account the arguments associated with the alternative position.

It may we worth noting that these findings are not in conflict with the conclusions of the previous chapter. I argued there that empirical findings show that our capacity for perspective-taking is sometimes involved in moral reasoning and moral judgment. This is consistent with it often not being used, or not being used appropriately.

2.2 Explaining egocentricity

Why do persons often not take alternative viewpoints properly into account? One possibility would be that they often discard other perspectives for one reason or another. While this surely happens sometimes, it does not seem to account for all of the above findings. It is unclear, for example, why people would first realise that others do not have access to certain privileged

[4] Indeed, not seldom do we in fact explain moral failings and moral conflicts in terms of agents not taking into account other viewpoints than their own or those of their group.

information and then discard this information when forming a social judgment. A more plausible explanation of most of the results is that people fail to perceive the perspectives of others accurately. They may either fail to see that others have an alternative viewpoint, or see that others have an alternative viewpoint but have an inaccurate understanding of it.[5] Our own subjective perceptions of events often feel perfectly objective, even when they are skewed by our own position. Because of this, we may often see no reason to even *consider* whether there are alternative viewpoints.

This explanation can be connected to the finding, discussed in the previous chapter, that perspective-taking requires cognitive resources such as attention and effort. Besides the experiment described in Chapter 3 (§2.1), there are additional studies that show egocentric biases to increase under time pressure and attention-demanding tasks. For example, as Caruso and Epley point out (2004), it has been found that when people evaluate their own skills in comparison with others they tend to concentrate on their own level of ability and consider the abilities of others insufficiently (Klar & Giladi, 1999). This leads them to overestimate their skill in activities in which absolute ability levels tend to be high, such as driving, and underestimate their skill in activities in which absolute levels tend to be low, such as chess. This egocentric bias increases when participants are under cognitive duress, such as when they have to hold a six-digit number in mind (Kruger, 1999). Such findings have been interpreted as showing that people have an 'egocentric default' when forming judgments the correcting of which requires cognitive resources (Epley et al., 2004; Epley & Caruso, 2004; Kruger, 1999).

In the previous chapter I have taken the fact that perspective-taking is cognitively effortful to imply that persons cannot apply the contract test when their resources are low. Drawing on the above findings, it may now be argued similarly that persons often have insufficient resources to identify and take into account alternative viewpoints with respect to moral principles, and thus insufficient resources to apply the contract test adequately. Indeed, given that participants in most of the studies considered do not appear to be under cognitive duress, such reasoning could be used to argue they may also not be able to apply it adequately under somewhat more favourable conditions than those of low cognitive resources.

That it is cognitively costly to consider other points of view is likely to be part of the explanation of the finding that people sometimes fail to take into

[5] I do not think the findings allow us to clearly distinguish between these two steps.

account alternative perspectives. But it cannot be the full story. Other studies show that people in similar circumstances can be moved to take into account relevant points of view. Indeed, studies have shown that both our tendency to rely on stereotypes and the egocentricity bias in our judgments can be decreased if people are *moved* to consider other viewpoints.

Galinsky & Moskowitz (2000) found that when persons are manipulated into adopting the perspectives of others, they are less likely to rely on stereotypical biases. Interestingly, perspective-taking did not only affect the conscious and explicit use of stereotypes but also the implicit reliance on them, as was shown by means of reaction-time experiments. In addition, the researchers found that perspective-taking can eliminate in-group favouritism. Participants performed a task in which they had to estimate how many dots were displayed on the screen. Irrespective of their responses, participants were told that their pattern of responses revealed that they were 'overestimators'. When they were subsequently asked to rate overestimators as well as members another group, the 'underestimators', on certain traits, participants rated members of their own group significantly more favourably than members of the other group (42.8 versus 52.8), in line with other experiments on the phenomenon of in-group favouritism. However, when participants were asked to write a short essay from the perspective of an underestimator before giving their rating, they rated underestimators and overestimators equally. Put differently, when participants engaged with the perspective of an out-group member, they did not develop an in-group bias. Daniel Batson & colleagues (1997b) report similar findings with respect to unfavourable attitudes towards stigmatised groups, such as victims of AIDS and the homeless: when persons were drawn to adopt the perspectives of individuals from these groups, their judgments of them became less negative.

Perspective-taking can also decrease the aforementioned egocentric bias. As I stated before, people tend to overestimate their own contributions to group tasks. In one study, Savitsky and colleagues (2005) asked participants to indicate how much they contributed to separate activities, such as writing or idea generation, that were part of a creative group project.[6] When the researchers summed these self-allocations per activity across members of the group, they were found to exceed the 100% significantly, indicating that individual members of the group tended to think they contributed more than the others. For example, the summed self-allocation for the writing process

[6] For another study with similar findings, see Caruso and colleagues (2006).

was 135.5% and that of idea generation was 147.8%. When participants were led to consider the contributions of other members before judging their own contribution, this overestimation decreased substantially. For instance, after considering other perspectives summed self-allocation for the writing process was 104.9%, and that of idea generation 110.4%.[7]

These findings suggest lack of cognitive resources is not the full story of why persons fail to take into account alternative points of view. Participants in these studies did not appear to have more cognitive resources—less distraction, more time—than participants in the studies discussed in the previous subsection, but nevertheless take into account alternative viewpoints and as a consequence judge more accurately. The difference is, Waytz and Epley (forthcoming) have suggested, that in these cases there is some 'trigger' that leads persons to engage in perspective-taking. In their words, "reasoning about other minds is more like rolling up a hill than like rolling down a hill. It requires a trigger to start and effort to maintain" (p. 25). Due to the perspective-taking manipulations, participants are directed to consider the perspectives of others and they subsequently come to take these alternative viewpoints into account.

In the above studies the trigger for engaging in perspective-taking comes from the experiments. Such external triggers may not be available when persons apply a contract test. There is also some evidence, however, that persons can trigger themselves into perspective-taking. Several studies have found that persons do take into account alternative viewpoints when doing so is instrumental to their ends. If people are offered financial incentives for forming an accurate social judgment, they are substantially more likely to consider appropriate points of view (Epley et al., 2004; Epley & Gilovich, 2005). The same occurs when they experience a loss of power or control (Waytz & Epley, forthcoming). The explanation for the latter finding, Waytz and Epley (forthcoming) argue, is that people who are not in control have more reason to understand the intentions of others. In line with this explanation, another study found that persons who experience power tend not to take an interest in alternative viewpoints (Galinsky, Magee, Ena Inesi, & Gruenfeld, 2006).

[7] Note that these findings do not fit the alternative explanation put forward in the previous section, that we would consider and identify other points of view and subsequently discard them. In that case the manipulation that had one consider other points of view should have had no effect.

Persons may also be motivated to engage in perspective-taking out of a need for social contact. When persons are in need of friends, they are more attentive to the emotions of others and more accurate in describing them (Pickett, Gardner, & Knowles, 2004). Lonely persons are more likely to attribute mental states to pet animals and even electronic gadgets than people who are not lonely, as are people who are experimentally induced to experience a sense of loneliness (Epley, Waytz, Akalis, & Cacioppo, 2008). In contrast, when persons feel socially satisfied they are less likely to consider the perspectives of others. In a series of experiments, Waytz and Epley (2012) found that persons who are led to feel socially connected are less likely to consider the minds of others and attribute mental states to them, in particular with regard to distant others. In order to have more insight into the relation between such 'dehumanization', as the researchers call it, and behaviour, they asked participants about their attitudes towards employing harsh interrogation techniques (i.e. torture) on individuals detained for plotting the September 11, 2001 attacks. Somewhat disturbingly, it was found that participants were more likely to dehumanize the detainees and endorse harming them when there was a friend of theirs in the same room, who was answering the questions independently, than when their was a stranger in the room. Waytz and Epley suggest that by being socially satisfied, persons did not consider the perspective of detainees.

While these findings do not support a very optimistic view of human psychology, they do show persons can move themselves to consider alternative viewpoints when they take themselves to have reason to do so. They therefore support the idea that we have a capacity for identifying alternative viewpoints, which is what the Practicability Assumption requires.

2.3 Identifying all relevant standpoints

That considering alternative viewpoints requires effort and motivation does, however, generate a concern: namely that we may, when applying the contract test, sometimes not be able to identify *all* relevant viewpoints. Scanlon shows himself to be sensitive to this problem, and proposes a solution: "we cannot envisage the reactions of every actual person. We can consider only representative cases" (p. 171). These representative cases are what Scanlon calls generic standpoints. To remind you, generic standpoints are the various roles in which persons are placed by a principle and from which they may as such object. I shall follow Scanlon in assuming that applying the contract test

requires not the consideration of every individual perspective but only generic standpoints.[8]

While the introduction of generic standpoints should lead to a serious reduction of the number of standpoints one must consider, it is does not remove the concern completely. It may still be difficult to identify all relevant generic standpoints. First, as I have explained before (2§4), identifying generic standpoints towards a given principle typically requires engaging with the perspectives of particular others. This may include the perspectives of persons who turn out to have the same generic standpoint. Second, the number of relevant standpoints may still be large. There may be different types of agents, different types of victims, and different types of bystanders who all hold different standpoints with respect to a principle in question.

Contract theorists may also have a response to these two problems. As they concentrate on unanimous agreement, there need be only one standpoint from which general acceptance or rejection of a principle is unacceptable in order for a principle to fail to be justified. A person applying the test may thus stop identifying standpoints once having found one from which it seems unacceptable. Furthermore, it seems plausible that we are sensitive to these standpoints. Take for example familiar principles such as the principle that promises made freely must be kept or the principle that one must help another person when doing so comes at little cost to oneself. When considering whether principles such as these may be overturned in particular cases, we are naturally directed to the standpoint of a person to whom a promise has been made or, in the second case, to that of a person needing help.

This response is, however, less effective to a third problem. The studies in the above subsection reveal that persons can take into account, and therefore identify, perspectives of certain salient others. In most studies, the relevant perspectives are of others whom participants are explicitly asked about. This suggests that, when applying the contract test, persons are able to identify standpoints associated with certain salient others, such as particular others

[8] A focus on generic standpoints fits Rawls's approach. The position behind the veil of ignorance is generic in the sense that we can all occupy it. In addition, Rawls has occupants of this perspective consider generic roles in society that they may have, rather than particular viewpoints of others. Determining the content of the agreement made in the original position thus requires a consideration of generic standpoints (see also footnote 18 of Chapter 2). It is less clear whether a focus on generic standpoints is consistent Gauthier's theory. Although the inhabitants of Gauthier's agreements are representatives of ourselves in the sense that they are idealised versions of ourselves, as far as I am aware Gauthier does not take the number of bargainers to be smaller than the number of represented persons. Indeed, this may not be consistent with his aim of deriving a contract that is in each individual's interest to accept.

who would be affected by an action under consideration. However, these may often not cover all the standpoints that should be taken into consideration. First, clearly, individual actions may affect others who are not part of one's own present situation. I will say that such persons occupy a *distant* standpoint.

There is a second way in which standpoints may be distant. As mentioned before (2§4), when evaluating a principle we must consider the implications of it being accepted as a principle for the general regulation of behaviour. In Scanlon's (1998) words, "when we are considering the acceptability of rejectability of a principle, we must take into account not only the consequences of particular actions, but also the consequences of general performance or nonperformance of such actions and of the other implications (for both agents and others) of having agents be licensed and directed to think in the way that the principle requires" (p. 203). For this reason, it is not sufficient to consider the perspectives of persons who would be affected by the action under consideration. While they certainly occupy relevant generic standpoints, there may be implications of general permission that go beyond the effects of particular actions. As such standpoints are not covered by the perspectives of particular persons in one's situation, I will again call them distant.

It may be helpful to give some examples of this. First, consider the action of crossing the speed limit for the thrill of it. Even though individual instances of this type of action may typically not have direct consequences for a salient other, the action being generally permitted would certainly lead to the deaths of some. Persons to whom this would happen occupy a standpoint that we must take into account when considering principles that permit crossing the speed limit. For a second example, consider a principle that allows one not to give aid to persons living in extreme poverty. While there may not be persons whose situations would be evidently worsened by the particular action (or rather, omission) covered by this principle, there are certainly others who would be worse off were the principle generally accepted rather than rejected.

As the previously discussed studies do not concern distant standpoints, they do not show we are able to identify them. To the contrary, there is reason to think such standpoints are in danger of being overlooked. We saw that identifying perspectives of persons who are in plain sight already requires significant effort and attention. Considering viewpoints of others who are not in one's present situation may be expected to require additional thought and imagination, and thus additional cognitive resources.

Moreover, the existence of distant viewpoints brings to the fore another requirement that must be satisfied in order to identify alternative standpoints towards a principle: to see that others would be affected by a type of action being permitted, one must have sufficient *information* both about the effects of actions and about others and their situations. For example, in order to identify the standpoint of third-world factory workers of a given company with regard to principles concerning the permissibility of buying products of this company, I must have information both about their working conditions and about how the choices of consumers affect these. Or to return to an example given above, to recognise that weaker road users such as cyclists hold a distinct standpoint with regard to the principle that allows persons to transgress the speed limit for the thrill of it, I must not only know that transgressing the speed limit increases the risk of road accidents in general, but in addition realise that due to their situation weaker road users would be significantly less safe under such a policy. As persons who have distant standpoints are per definition less closely related to one's own present situation, one is more likely to be ill informed about them and as a consequence overlook their standpoint.

2.4 Conclusions

The findings presented in this section reveal several things about our capacity to identify relevant other points of view. They show first of all that we are not very sensitive to the perspectives of others. Identifying other points of view does not work automatically, but requires attention, effort and motivation. The fact that people can be moved to consider other perspectives when it is in their interest to do so indicates, however, that they can consider relevant other viewpoints when they take themselves to have reason to do so. It seems that they understand, at least implicitly, that there is a trade-off between effort and accuracy, and often sacrifice accuracy in favour of ease.

This supports the idea that we have a capacity to identify alternative viewpoints, which is what the Practicability Assumption requires. But the findings also show this capacity to be constrained in ways that may affect our performance when applying the contract test. Identifying alternative viewpoints towards a principle requires effort, motivation, and information. When one of these three resources is insufficiently available, relevant standpoints may be overlooked which in turn may lead to mistaken conclusions about the justification of principles. I return to this in sections 4 and 5, after first considering our ability to understand identified alternative standpoints.

3 Understanding alternative standpoints

After having identified alternative standpoints, the second step in the contract test is to determine what objections to the principle in question being either accepted or rejected may be posed from them. Moral contract theorists hold different views about the grounds for objections. Scanlon holds that objections depend on how acceptance or rejection of the principle would affect aspects of a person's life that she has reason to care about, including her well-being, liberty, opportunities, aims, relations, control over her body, and view of herself.[9] Gauthier, on the other hand, holds that a person's reasons for objecting depends on the extent to which acceptance or rejection of the principle satisfies her considered preferences.[10] Despite this difference, both varieties of the contract test require that, in order to determine potential objections, one must first and foremost understand what implications acceptance or rejection of the principle would have for them. This second step of the contract test thus requires one to engage with alternative standpoints towards the principle in question and determine what acceptance or rejection of the principle in question would mean for them.

Just as the previous section discussed what findings on perspective-taking show about our ability to identify who would be affected by a given principle, the present section discusses what they show about our ability to identify how they will be affected. Given the close relation between identifying a point of view and understanding it, there is overlap in the relevant empirical findings. Indeed, many of the findings presented in the previous section can be interpreted as either showing that persons often fail to identify relevant perspectives or that their understanding of these perspectives is limited. Not surprisingly, then, I will in this section also conclude that we often only have a limited grasp of alternative perspectives. In order to find out to what this shows about our ability for perspective-taking, I will again consider how the findings should be explained.

3.1 Egocentric interpretations

We have already seen that people have some accuracy in understanding other points of view. This is shown, for example, by the fact that older children and adults usually succeed in the false-belief task by predicting the behaviour of agents with false beliefs correctly. That the use of inaccurate stereotypes

[9] See Scanlon (1998, pp. 203-204 and 214-215; and also 2003, pp. 182-183)
[10] See Gauthier (1986, in particular Chapter 2).

decreases when people consider the perspectives of others also supports this. But although we evidently have some grasp of other points of view that we consider, other studies suggest our understanding of them is not very accurate. I will examine two lines of research. The first line of research concentrates on determining the accuracy of our interpretations of others. The second and larger line of research has focused on identifying biases in our thinking about others.

In a series of experiments, William Ickes and colleagues have tried to establish how accurate our interpretations of others are—what our 'empathic accuracy' is (Ickes, Stinson, Bissonnette, & Garcia, 1990; Stinson & Ickes, 1992). Ickes and colleagues' methodology comes down to comparing the thoughts and feelings that perceivers attribute to target persons with the thoughts and feelings that these target persons report to have (for a discussion, see Hodges, 2008). To do so, they videotape target persons while they are talking to either a camera or another person. Target persons are subsequently asked to watch this videotape, and instructed to stop the tape whenever they remembered having a specific thought or feeling during the conversation and write the contents of these thoughts and feelings down. Ickes and colleagues let other participants watch these tapes and judge what the videotaped target persons were thinking or feeling. When participants watch the video, it stops at each moment that the target person stopped the tape and wrote down his thoughts and feelings, and participants are asked to write down what they think the target person is feeling or thinking. A series of experiments done with this methodology has found strangers to be correct in estimating another's thoughts and feelings around 20% of the time. Friends have an empathic accuracy of about 30%. Although this confirms that we have some insight into the minds of others, it clearly is not very much.

The second line of research concerns the biases that are in play when we think about others. Social psychologists have studied these extensively. One of these biases played a role in the previous section. Due to experiencing the world solely first from our own point of view, we often fail to see that others may have a different appreciation of events than we do. However, our egocentricity goes further than just this. Abundant studies have shown that even when we recognise that others have a different point of view to ours, our interpretation of it tends to be biased by our own point of view.

First of all, there is extensive evidence for something that has been called the 'false consensus' phenomenon. Over a hundred studies have revealed that people tend to overestimate the extent to which others share their own

personal characteristics— preferences, expectations, fears, habits, etcetera (Mullen et al., 1985). In one of the earliest studies on this phenomenon, Katz and Allport (1931) found that students who admitted to cheating in tests tended to overestimate how many of their peers cheated. In another now classic study, Ross, Green and House (1977), found that whether participants expected that others were willing to walk on campus wearing a sandwich board (either stating 'Eat at Joe' or 'Repent!') depended on their own choice to participate. While participants who were willing to do so estimated that 65% would wear the board, participants who were not willing estimated that only 31% would do so. Several studies have shown that the extent to which people expect others to agree with certain statements depends on how strongly they agree with these themselves (Krueger & Clement, 1994; Ross et al., 1977). To give some examples, with regard to the statement "I like poetry", people who themselves endorsed the statement estimated 56% of the population would agree with it, while people who did not endorse it estimated 48% would agree with it; with regard to the statement "I seldom worry about my health", endorsers estimated 45% of agreement while non-endorsers estimated only 34% of agreement; and with regard to the statement "I have no fear of spiders", endorsers estimated 51% would agree with it while non-endorsers estimated it would be 37%.

Besides a 'false consensus' there is also evidence for a 'curse of knowledge'. Several studies have found that when we have privileged knowledge about something we overestimate the extent to which others have this knowledge as well (for an overview see Nickerson, 1999). For example, one study found that participants who observed a negotiation and were told about the motives of the negotiators (e.g., to be accommodating or to be assertive) overestimated how clear that motive would be to the other negotiator (Vorauer & Claude, 1998). In another study, participants who were informed that a certain protagonist was not amused by a stand-up comedian overestimated how likely an uninformed person would be to identify that the protagonist was being sarcastic when he said it was 'hilarious' in a voicemail message to that person (Epley et al., 2004). Yet another study found that participants who were asked whether they could taste the difference between Pepsi and Coca Cola, while informed of the drinks' identities, overestimated the extent to which individuals who would not be informed about them could identify them correctly (Epley et al., 2004).

The false consensus and the curse of knowledge effect are usually taken to show that our interpretations of others tend to be egocentrically biased.

Kawada and colleagues (Kawada et al., 2004) have shown this is also the case with goals. In one study, they found that what type of goals persons tend to ascribe to others depends on what type of goals they tend to set themselves. Drawing on the finding that people differ in how they are disposed regarding so-called achievement situations (Chiu, Hong, & Dweck, 1997), the researchers first used a questionnaire to distinguish between, on the one hand, persons who are disposed to aim to learn new things when in situations in which there is something to achieve on the one hand, and, on the other hand, persons disposed to aim to show what they can already do when in such situations. They found that when persons of the first type observe others in similar situations they tend to ascribe learning goals to them as well, whereas persons of the second type are more likely to interpret others as having performance goals. In another study, Kawada and colleagues found that participants who possessed a goal to compete with others attributed more competitiveness to characters involved in a Prisoner's Dilemma than participants who did not have this goal. Interestingly, this occurred both when participants were explicitly told to be competitive and when they were primed to be competitive by means of a seemingly unrelated task. Apparently, we not only unconsciously project characteristics of ourselves onto others, but also unconsciously project characteristics we are not conscious that we have.

Van Boven and colleagues (2000) have shown that our understanding of other points of view is also influenced by more affective aspects of one's own situation. It is a well-known finding in social psychology that simply by owning an object people tend to value it more. Van Boven, Dunning and Loewenstein (Van Boven et al., 2000) found that individuals do not account for this in their expectations of others: both owners and buyers overestimate similarity between how much they value a commodity and how much it is valued by people in the other role. This may lead to suboptimal behaviour in settings with economic consequences (Van Boven, Loewenstein, & Dunning, 2003).

In another study, Van Boven (2005) and colleagues asked participants to predict how much money a randomly selected student would have to be paid in order to dance alone on a stage in front of a full auditorium to the song 'Super Freak' by Rick James. Participants predicted such a person would be willing to do it for $13 dollars, whereas students who actually faced the prospect of dancing asked substantially more ($53). Predictors also underestimated the extent to which potential dancers would be concerned with what others would think of them in comparison with how much they would be concerned with the

money they could gain. Studies such as these two indicate that our interpretations of others are often insufficiently sensitive to affects associated with their situation.

That our actual situation affects our understanding of others is perhaps best illustrated by a study in which Van Boven and Loewenstein (2003) interviewed individuals either before or after engaging in a vigorous cardiovascular activity. Participants read a description of three hikers who were lost in the Colorado mountains without food or water and were asked to predict which experience would be more unpleasant for the hikers, hunger or thirst. Participants who were asked before the exercise answered thirst 57% of the time, while those who were asked after the exercise gave this answer 88% of the time. Such a finding indicates that the accuracy of one's interpretations of other perspectives may be affected by visceral aspects of one's situation, for good or for ill.

In order to find out what these findings mean for the plausibility of the Practicability Assumption we need to know why they occur. I turn to this now.

3.2 Explaining egocentric interpretations

Many psychologists interpret the finding that our interpretations of others are egocentrically coloured as showing that we use our own point of view to understand those of others (Epley et al., 2004; Kawada et al., 2004; Nickerson, 1998).[11] In the words of Epley and colleagues, "people adopt others' perspectives by initially anchoring on their own perspective and only subsequently, serially, and effortfully accounting for differences between themselves and others until a plausible estimate is reached" (p. 328)[12]. There are additional indications for this.

First, the above biases have been found to occur not only when people form judgments about others but also when they form judgments about themselves in other situations (Van Boven & Loewenstein, 2005). For example, just as persons overestimate how able uninformed others are at distinguishing Coca Cola from Pepsi when informed about this themselves, they overestimate how able they themselves would be were they not informed

[11] This is of course not the only explanation, but it is the dominant one. For an alternative explanation consider Karniol (2003).

[12] Philosophers of mind are likely to recognise that this fits nicely with the simulation theory of mindreading that has been defended most notably by Robert Gordon (1986; 2009) and Alvin Goldman (1989; 2006).

(Epley et al., 2004). Similarly, just as people underestimate how much money others need to be given to dance to 'Super Freak', the number they give when asked how much they themselves would need to be given is much lower than the $53 that people in fact demand. In addition, for at least some of these findings, statistical analyses have been used to demonstrate that people made judgments about others by first predicting how they themselves would think or feel in the other's situation (Van Boven & Loewenstein, 2005). On the basis of such considerations, Van Boven and Loewenstein (2005) write that "the accuracy of social predictions depends critically on the accuracy of self-predictions" (p. 47).

Second, as was mentioned before, egocentricity has been found to increase when cognitive resources are low. For example, when participants who knew that a message was intended sarcastically were asked whether uninformed participants would judge it as such under time pressure, 66% said it would, while participants who were told to 'take their time' said so only 50% of the time (Epley et al., 2004). This supports the idea that when interpreting another perspective persons starts from an egocentric default and make effortful adjustments to account for differences between themselves and an occupant of that perspective.

If we indeed interpret alternative points of view by making adjustments to our own, the studies in the previous section show that we often do not make *sufficient* adjustments when interpreting others. As in the previous section, whether this has implications for the plausibility of the Practicability Assumption depends on whether something can be done about it. It must therefore be asked why people tend not to make sufficient adjustments.

Three causes may be distinguished, two of which we have seen before. The first, which has been emphasised by social psychologists, is that accuracy is costly (Epley, 2004; Kawada et al., 2004). Self-referent knowledge is more readily available than knowledge about others. As making adjustments comes at the cost of effort and time, people tend to terminate adjusting once a plausible estimate of another's point of view is reached. This fits well with the finding that when participants are motivated to be accurate through financial incentives, accuracy increases (Engelmann & Strobel, 2000; Epley et al., 2004; Epley & Caruso, 2004). Epley, Keysar, Van Boven and Gilovich (2004) studied this more directly. As with the experiment discussed in Chapter 3 (§2.1), participants were either led to believe that a given message was sarcastic or that it was sincere. Participants were then asked to judge how likely others were to believe the message to be sincere rather than sarcastic, in either of two

ways. A first group of participants was asked to estimate the number of others that would interpret the message as sincere rather than sarcastic. A second group of participants was not asked to give one estimate of this number, but instead to give the *range* of values in which they expected this number to be. Epley and colleagues found an interesting relation between these estimates and ranges. Of the participants who were led to believe the message was sincere, the estimate provided by those in the first group did not differ significantly from the upper bound of the range given by those of the second group. Put differently, when they believed the message to be sincere, participants in the first group gave an estimate that was skewed strongly to the sincere end of the range of values given by the second group. In contrast, of the participants who were led to believe the message was sarcastic, the estimate provided by the first group did not differ from the *lower* bound given by the second group. Stated differently, when they believed the message was sarcastic, participants in the first group gave an estimate that was strongly skewed to the sarcastic end of the range of values considered plausible by the second group. Apparently, Epley and colleagues conclude, participants who were asked to give only an estimate took their own point of view as a starting point, and stopped adjusting once the first plausible value was reached. Epley and colleagues interpret this as evidence for their view that when considering an alternative point of view, persons engage in a "process of adjustment from an egocentric anchor that terminates once a plausible estimate is reached" (p. 335).

This first cause of inaccuracy is similar to the explanation given in the previous section for why people often do not consider other salient points of view. It is therefore also open to a similar response: people should simply try harder. Additional effort will only increase accuracy up to a point, however, as there are two other causes for why we make insufficient adjustments.

The second cause is lack of information. In order to understand another person's point of view one must have sufficient information about that person and his situation. In terms of the above account of perspective-taking, one must be aware of differences between one's own situation and that of the person whose perspective one seeks to understand, so that one can make the proper adjustments. If I fail to observe that Maxi is not around when Mother replaces his chocolate from the box to the cupboard, I will not attribute the (false) belief to him that the chocolate is in the box and will subsequently mistakenly predict that he will look in the box. If I am unaware that the curtains in my friend's new house have been put up by herself rather than the

previous owner, I will fail to understand that telling her they are ugly would hurt her feelings. If I do not know that my upstairs neighbour went to bed early, I will fail to see that ringing her doorbell after 21h would awaken her from her sleep and annoy her. And if I do not know that my colleague has a strong dislike for Hollywood movies, I may mistakenly think that the latest James Cameron blockbuster is an appropriate topic for conversation. Such examples illustrate that whether we understand another's perspective depends for an important part on how well informed we are about her and her situation, including characteristics such as her beliefs, habits, and preferences. We need such information in order to know what adjustments to make.

Given their mundane nature, the examples also illustrate that we are often *not* sufficiently informed about another's situation to do so. We typically have only scarce individuating information. Fortunately, perspective-taking can also be facilitated by information of a more general type such as stereotypes. Such psychological generalisations play a role in the examples just given. We know that when people observe events they tend to remember them, that direct criticism tends to hurt feelings, that people do not like to be woken up for no good reason, and that many people like to watch Hollywood blockbusters. Information about another's situation and psychological generalisations often work in tandem: when learning something about another's situation, we can apply a generalisation, and thus increase our understanding of that person's point of view. However, as is well known from the psychological literature on stereotypes, psychological generalisations are far from flawless. While most people like Hollywood movies, certainly not everyone does. Psychological generalisations may thus sometimes decrease our understanding of another individual's point of view.

We seldom have *all* the relevant information when attempting to understand another's point of view on something. A person's perspective may depend on subtle details of his situation and himself that are difficult to observe, including idiosyncratic beliefs and desires. Furthermore, our knowledge of relevant psychological generalisations is limited. This can also be seen in the above studies on egocentricity. In each of them, participants answered questions about others in situations about which they had only limited information. Take for example the studies on the false consensus effect. Surely, people are not aware of the exact proportion of members of the population who like poetry or who are afraid of spiders.

Given limited information about others, it is not surprising that persons tend to rely on their own point of view to understand those of others. It may

often be the best available method to make sense of others. Nor is it surprising that people tend not to make sufficient adjustments, as knowing what adjustments to make requires detailed information about differences between oneself and the other. Lack of information thus explains not only why our interpretations of other points of view tend not to be very accurate, but is also part of the explanation of why they tend to be egocentrically biased. The finding that egocentric biases may decrease when people are better informed supports this. Several studies have found that the false consensus effect decreases substantially when people are better informed, for instance about the views of particular third-persons (Clement & Krueger, 2000; Engelmann & Strobel, 2000). That is not to say, however, that the egocentric bias would disappear if people are both motivated and fully informed, as there is another factor.

The third cause for making insufficient adjustments is that our own perspective constrains our interpretations of other points of view. Necessarily, when we consider an alternative point of view regarding a given object, we have already construed the object from our own point of view. Having a correct interpretation of another perspective requires not just taking into account aspects of another's situation, but also inhibiting those influences of one's own perspective that do not apply to the other. This includes traits, beliefs, goals, and desires, but also, as we saw in the final part of the previous section, emotions and physiological arousal.

That we may not be conscious about such characteristics nor about their influence on our interpretations of other viewpoints is one reason why it may not always be possible to inhibit their influence. However, it is unlikely that being completely informed and conscious about this would fully remove this third cause of egocentricity. People may have difficulty correcting for aspects of their own point of view even when they are informed these do not apply to a target perspective. Take for example the previously mentioned studies in which participants had to follow the instructions of a speaker who could not see the same objects as they did (Epley, 2004; Keysar et al., 2003). Even though participants were aware that the speaker did not know the identity of an object that they were carrying in a bag, they regularly interpreted descriptions of a similar object visible to both of them to refer to this hidden object. On the basis of such findings, Epley concludes that "considering another's perspective cannot alter one's construal of an event any more than actively trying to see colorblind will render a person unable to distinguish red

from green" (p. 1467). A certain degree of egocentricity is inherent to our ability for perspective-taking.

3.3 Conclusions

What does the above show about our ability to understand other points of view? Let's first briefly summarise the findings. Empirical findings on perspective-taking suggest that while we certainly tend to have some understanding of perspectives that we consider, this understanding is far from excellent. That we tend to have some understanding of them is not only shown by the fact that we succeed in perspective-taking tasks and have some empathic accuracy, but also by the fact that our egocentrically biased judgments are in the right direction. For example, while participants who believe a message to be sarcastic overestimate how likely others are to recognise it as such, the findings also clearly show these same participants to realise that others may fail to recognise it as sarcastic. Similarly, studies on the false consensus effect find that there is a positive association between our predictions and actual agreement, even though we tend to overestimate the extent to which others agree (Krueger & Clement, 1994). That this understanding is not excellent is shown by the same studies. Our empathic accuracy is not very high, and our interpretations of other points of view tend to be biased by our own present situation.

What does this mean for our ability to apply the contract test, and in particular for our ability to understand the objections that may be posed from alternative standpoints regarding moral principles? That we tend to have a certain degree of understanding of perspectives we consider suggests we are able to reach some understanding of the objections that may be raised from alternative standpoints. However, the above also provides reason for thinking that our ability to recognise such objections is constrained in certain ways. First, understanding an alternative point of view requires cognitive resources and information about differences between one's own position and that of occupants of the standpoint. Second, even when these resources are available, we may not be able to suppress our own point of view regarding an action or principle when interpreting the alternative viewpoint, which would result in an egocentrically biased interpretation. The next section considers in detail how these limitations, as well as those identified in the previous section, may be reflected in our performance when applying the contract test.

4 Perspective-taking inaccuracy and the contract test

The previous two sections discussed empirical findings regarding our accuracy in perspective-taking to find out how able we are at engaging in the kind of perspective-taking required for the contract test. Such findings confirm that we can to a certain extent identify and understand alternative points of view, I argued. However, they also reveal that our perspective-taking accuracy is limited by both the availability of cognitive resources and information. Furthermore, there is reason to think that our interpretations of other points of view are often coloured by our own present perspective. I have in the previous sections already mentioned various ways in which the resulting inaccuracy may be reflected in our performance with the contract test. In the present section I will discuss this more systematically, distinguishing several errors that may occur. The next section discusses what such errors mean for the question of whether we can apply the contract test adequately or not.

Applying a contract test to an action involves determining whether the action is in conformity with principles for the general regulation of behaviour that everyone has reason to agree to. As before, several steps may be distinguished. Besides identifying and understanding standpoints implicated by the general acceptance or rejection of a principle that allows a certain type of action that we wish to evaluate, I will now also add the prior step of singling out the appropriate principles. As such principles will, when stated explicitly, include references to persons in various relations towards the action and the implications that allowance or disallowance of the action would have for them, perspective-taking is required for selecting the appropriate principle.

I shall thus distinguish three steps involved in applying the contract test to an action: (1) singling out a principle that would allow acting in that way under those circumstances, (2) identifying alternative standpoints with regard to the principle, and (3) determining what objections may be posed from these standpoints. I shall argue that limited perspective-taking accuracy may lead to errors in each of the steps.

Before starting the discussion, two things should be noted about this analysis of the contract test. First, the steps of the contract test overlap with one another, at least when they are performed correctly. To identify correctly which principle would permit one's action, one must already be well informed about the burdens that the principle imposes on certain positions. Similarly, in order to identify standpoints implicated by a principle, one must be aware of

what burdens the principle imposes on others. Nevertheless, for purposes of exposition it is helpful to discuss the steps separately.

Second, these steps do not constitute the full test. Besides determining what objections may be made from various standpoints given the implications that a principle has for them, one must assign weights to these objections: how *strong* a reason do the implications give persons occupying those standpoints for wanting the principle to be accepted or rejected? As I mentioned before, different contract theorists hold different views about how such normative questions must be addressed. Furthermore, in order to decide whether a principle would be accepted, one must consider what alternatives there are. For example, an alternative to a principle that allows a certain type of action would be a principle that *dis*allows such actions (Scanlon, 1998, p. 195). Note that errors that may occur in the above three steps may just as well occur when comparing alternatives.

4.1 Step 1: discerning the appropriate principle

To evaluate an action by means of the contract test, one must first identify a principle for the general regulation of behaviour that would allow the action. Limited accuracy in perspective-taking can lead one to fail to single out the appropriate principle for the situation at hand.

Which principles allow a given action depends on the nature of the action, and in particular the action's implications for others. Say another person asks for my assistance, and I am considering not doing so because I am already involved in some activity. What principle would permit me to do this depends, among other things, on the degree of inconvenience involved in giving assistance and on the importance of receiving assistance. A principle allowing one to refrain from saving another from great harm when doing so would involve little cost for oneself is rather different from, say, a principle that allows one to refrain from providing a small benefit when it would prevent one from engaging in a valuable project. In order to identify a principle that would allow the action under consideration, I need thus to compare how burdened I would be by helping and how burdened others would be if I refrained from helping. Clearly, perspective-taking inaccuracy can lead to mistakes in this first step. If I fail to recognise others who would be affected by my action or misinterpret the implications that my action would have for others, I may end up with a mistaken belief about what principle would allow my action.

4.2 Step 2: identifying alternative standpoints

Having identified a principle that would permit the action under consideration, one must determine whether it is a principle that everyone has reason to agree to. One must therefore consider alternative standpoints that can be occupied regarding the principle besides the one presently occupied by oneself. As considering every individual's perspective is not practically possible, I follow Scanlon's suggestion that it is sufficient to consider generic standpoints.[13] With regard to a principle allowing one to refrain from giving assistance under certain conditions, two of the relevant standpoints are that of a person who would be giving the assistance and that of a person who would be receiving the assistance.

The mistake that perspective-taking inaccuracy may generate in this second step has already been discussed extensively in section 2: persons may fail to identify relevant standpoints towards the principle that they consider. All three causes of perspective-taking inaccuracy may lead to this mistake. A person applying the contract test may not have sufficient cognitive resources to consider alternative points of view. Identifying relevant standpoints regarding a principle requires one to think through the implications of that principle being accepted for the general regulation of behaviour. One needs thus to imagine in what situations persons would be placed were a principle generally accepted, given the various properties that they presently have. Such a thought experiment is effortful and difficult, and a person may thus fail to take into account all relevant points of view.

A person applying the contract test may also fail to identify alternative standpoints towards a principle due to lack of information. To identify a given standpoint, one must be aware that persons occupying this standpoint would be affected by the principle. As I mentioned before (§2.3), this requires information about the effects of actions permitted by the principle as well as information about the situations of those who would be affected. While a person applying the contract test does not need to know the particular others who would be affected, she must know sufficient about their situation to recognise *that* they would be affected by the principle and therefore occupy a relevant standpoint.

---

[13] As I mentioned in footnote 8 of this chapter, due to his focus on individual preferences this move may not be acceptable for Gauthier. In that case, applying his contract test comes with an additional complexity, as it does not seem possible for us to consider everyone's perspective. The various errors discussed below do also apply to his contract test, however.

By concentrating on generic standpoints rather than specific perspectives, contract theorists reduce the amount of information that is required for this step of the contract test. Nevertheless, in order to identify particular standpoints as distinct from other standpoints one may still require information about details of the situations of others. As I mentioned before (2§4), despite being generic, standpoints may be quite specific. A given principle may affect distinct individuals in very different ways, depending on their capabilities, aims, and the conditions in which they are placed, among other things. For example, a principle that permits the rich to give only minimal aid to those less well-off will have very different implications for an individual who is not very well-off and and an individual who lives in extreme poverty. These individuals therefore occupy different generic standpoints with regard to the principle. (For some other examples consider 2§4.)

Finally, a person applying the contract test may fail to identify standpoints by being too immersed in her own point of view. Studies in the previous sections show that we tend to project characteristics of ourselves to others when considering their perspectives, and therefore tend to form an egocentric interpretation of their viewpoints. When applying the contract test, such projection may lead a person to fail to recognise alternative perspectives towards a principle. For example, when a person who enjoys driving faster than the speed limit considers other points of view regarding a principle that allows such behaviour, projection of this property may lead her to overlook the particularly relevant standpoint of persons who do not enjoy driving faster than 120 km/h.

4.3 Step 3: understanding the implications

The third step of the contract test is to determine what objections may be posed against a principle from the affected standpoints. Even more so than for the previous step, this requires one to be well informed about their situation. In which way acceptance of a principle affects a person's security, liberty, opportunities, projects, relations, self-respect, and so forth, depends on how that person is situated with regard to the principle in question. As with the previous step, one must not just be attentive to the effects of performance or nonperformance of individual actions but to the implications of general performance or nonperformance of such actions. Clearly, effects of widespread performance can be very different from the effects of individual instances. But, as Scanlon (1998, p. 203) points out, general acceptance of a principle may also have less evident implications for either agents or others. Think of a principle

that disallows persons to prevent injury to a family member if they can prevent a much greater injury to a stranger. Acceptance of such a principle would surely place significant psychological burdens on potential agents and may affect their lives beyond the situations to which it applies.

Due to limited perspective-taking accuracy, persons may fail to recognise implications of a principle. Say I am listening to some rather loud music and suddenly wonder whether it is permissible for me to have the music turned to this volume. Applying the contract test to my situation, I consider whether a principle that allows me to listen to loud music in the evening is one that others may reject, and identify my neighbours as occupying a relevant standpoint with regard to this issue. When attempting to determine the implications for them, all three causes of perspective-taking may lead me to form an inaccurate judgment. Immersed in my present situation, being somewhat tired and distracted, and unaware of the exact specifics of the situation of my neighbours, I may fail to to appreciate exactly how loud and thus how disturbing my music is for my neighbours. In that case, my judgment about the objections they may pose will be inaccurate as well.

To what extent a principle burdens someone may depend on talents and capabilities, needs and vulnerabilities, on the social and economic conditions in which persons are placed, on their cultural norms and religious beliefs, and on their projects, aims, and tastes. While my neighbours share a similar standpoint in that they are forced to listen to my music, they may thus be differently burdened due to such individual differences. A neighbour who enjoys spending her evenings reading philosophy, or a neighbour who goes to bed early because she needs much sleep, will be affected differently by a principle that permits loud music in the evening than a neighbour who spends his evenings watching television and has a later bedtime.

Findings presented in the previous section suggest that the larger the differences between a person's own perspective and a target standpoint, the more inaccurate her interpretation of that standpoint will be. Persons are therefore particularly likely to fail to recognise implications for another when they would, due to having different characteristics, not endure such implications had they been in the other's situation themselves. If I do not share my neighbour's passion for reading philosophy in the evenings, but instead am someone who spends his nights listening to loud music, simply imagining myself in their situation will lead to a flawed perception of the implications she suffers. To reach an accurate understanding, additional adjustments need to be made to my own point of view, depending on how much I differ from them.

Given that persons will usually lack information about the extent to which such differences obtain, that making such adjustments is costly, and that our interpretations are coloured by our own point of view, they are likely to adjust insufficiently.

To sum up, limited perspective-taking accuracy can lead to mistakes several points when applying the contract test. It can lead one to fail to discern the principle appropriate to one's situation, overlook standpoints implicated by a principle, and form mistaken judgments about the objections that may be posed against the principle from these standpoints. Such errors may in turn lead one to draw mistaken conclusions about whether an action does or does not satisfy the test.

5 Perspective-taking inaccuracy and the contract test's correct-usability

The above discussion reveals that perspective-taking inaccuracy may lead to several types of errors when applying the contract test. Whether this poses a problem for the assumption that we can come to use the contract test adequately as an instrument for moral justification depends on the answers to two questions. The first question is how often such mistakes will result in mistaken conclusions. This is the question to which I turn now. More precisely, I will consider whether actual persons, with the accuracy in perspective-taking that people in general have (and with the knowledge of others and their situations that people in general have), who apply the contract test individually, thus without communicating or collaborating with others, are likely to draw mistaken conclusions when applying the contract test. The second question is whether it lies within our power to reduce the likelihood of making such mistakes. Put differently, to what extent can we improve our performance with the contract test? This question will be addressed in the next chapter.

Chapter 2 introduced two criteria that should, to an appropriate extent, be satisfied for the contract test to be an adequate instrument of moral justification. The first criterion is determinacy: that when persons apply the contract test, it provides an answer. The second criterion is correct-usability: that when persons apply the test, they tend to do so correctly. I take the first question just mentioned to concern this second criterion: are we so likely to err when applying the contract test that it does not satisfy this criterion?

As I mentioned in that same chapter, I do not have an exact measure for whether this criterion is satisfied. I shall go about it differently, and consider whether there are important cases in which mistakes caused by limited perspective-taking accuracy are likely to lead to incorrect conclusions. If there is reason to think we cannot apply the contract test correctly in important types of cases, there is reason to doubt its general correct-usability.

When evaluating an action by means of the contract test, perspective-taking errors may lead one to draw mistaken conclusions about whether the action is in conformity with principles that would be the object of agreement. As we saw in the previous section, perspective-taking errors may lead one to select an inappropriate principle for the action under consideration, in the sense that the type of action is not covered by it. Or they may cause errors in later steps, leading one to mistakenly conclude that a principle would or would not be the object of agreement. Perspective-taking errors may thus lead one to conclude that an action is (dis)allowed by principles everyone has reason to accept, or vice versa, even though this is not the case.

I say 'may' as errors do not *have to* lead to a mistaken judgment about the action. A person may select an inappropriate principle or mistakenly think a principle would be the object of agreement due to such errors, yet conclude correctly about whether the action under consideration satisfies the contract test. This happens when there is some *other* principle to which the action confirms that *would* be the object of agreement. For example, I may conclude correctly in a particular case that lying to another person would not satisfy the contract test, even though I arrived at this conclusion by judging mistakenly that the principle that forbids lying in general would be the object of agreement.[14]

That perspective-taking errors do not need to result in mistaken conclusions about *principles* is even more obvious. A person may fail to identify certain objections to a principle, or even fail to identify a standpoint that would be implicated by it, and nevertheless draw a correct conclusion about its acceptability. This is particularly likely to occur when considering principles

[14] Such mistakes may be problematic in other ways. First, while the person's conclusion about the action may not be incorrect, it may nevertheless not be justified. This is particularly so when an action does or does not conform to the contract test because of other principles than the ones that were considered. My conclusion would merely be correct by accident. Second, when a person would rely on such conclusions about principles at a later point in time under circumstances that are somewhat different, there may no longer be a justified principle to which the action happens to conform. Put differently, even if a mistaken conclusion about a principle does not lead to a false judgment about an action at first, it may yet lead to such false judgments eventually.

that would impose substantial burdens on certain conspicuous standpoints were they accepted whereas their rejection imposes much smaller burdens on others, or the other way around. Many familiar moral principles are like this. Moral transgressions often have direct negative consequences for certain others that strongly outweigh the benefits provided to agents performing these transgressions. Think of stealing, making a false promise to get what you want, causing emotional injury to another in order to feel better about yourself, or not helping a person who is about to drown in a pond. Each of these cases is regulated by principles the acceptance of which imposes heavy burdens on certain victims, while agents are not heavily burdened by their rejection. As the burdens imposed on victims are very salient and outweigh other relevant considerations, a person applying the contract test to these principles may make all sorts of perspective-taking errors and yet conclude correctly that the principle would not be the object of agreement. There seem to be many cases in which persons may apply the contract test successfully even when their grasp of relevant alternative standpoints is limited at best.

However, besides such easy cases there are also what we may call *tricky cases*, in which perspective-taking inaccuracy *does* generate a serious risk of applying the contract test unsuccessfully. I shall describe three such types of cases. First, there is an increased risk for mistaken conclusions when considering a principle that is only unacceptable from what I referred to as distant standpoints (§2.3). When an agent evaluates a given action, there will usually be particular others who would be affected by the action. Their standpoints are unlikely to be overlooked. However, the implications of the action may stretch beyond salient others, affecting persons less close to the action at hand. Moreover, as I explained before, there may be persons who would be affected by general acceptance of the principle, even though they would not be affected by single instances. If persons in either of these situations are differently affected than the salient others, they occupy standpoints that are more likely to be overlooked due to perspective-taking inaccuracy. And if such a distant standpoint towards a principle under consideration is the only standpoint from which it is unacceptable, overlooking it will lead one to mistakenly conclude that the principle would be the object of agreement.

Second, there is an increased risk for mistaken conclusions when we evaluate principles only unacceptable to persons very different from ourselves. One reason why our limited accuracy in perspective-taking is unlikely to cause mistakes when we think about familiar moral principles is that the implications

imposed on victims would also be endured by ourselves, had we been in their situation. As the implications imposed on these standpoints do not depend on individual differences, we need to make relatively few adjustments to our own present point of view to reach a sufficient understanding of them. Many principles, however, do have different implications for persons with different characteristics. As I explained in §4.3, the more we differ from others with respect to characteristics that influence how they are affected by a principle, the more likely we are to have an inadequate interpretation of their standpoint with regard to the principle— or even to fail to take the standpoint into account altogether. Therefore, when we consider a principle that is only unacceptable by certain persons who, due to their characteristics, occupy a standpoint with regard to the principle that is far removed from our own present standpoint, there is an increased risk that we mistakenly conclude that the principle would be the object of agreement.

Third, there is an increased risk for mistaken conclusions when considering principles that impose substantial costs on some if accepted but that also impose substantial costs, either on the same individuals or others, if rejected. Think of a situation in which you can only help another by sacrificing a project that is important to you, or a situation in which another person can only be calmed down through a false promise. To arrive at correct conclusions in such cases, one must have an accurate understanding of the various costs and benefits that principles governing these situations would impose.

It is important to recognise that such difficult cases may even arise with regard to our most familiar moral principles, such as the principle that forbids lying. A principle that imposes a general prohibition to lie would not satisfy the contract test. The person that would be killed were you not allowed to lie to his nemesis has a standpoint from which the principle that forbids you to lie can reasonably be rejected. Less dramatically, say you can spare another person's feelings by lying about your appreciation of his new clothes. If we look at it from the perspective of a person who would be lied to under such circumstances, it seems that lying may cause less harm than telling the truth. Whether this is so depends on additional relevant factors about the situation, such as when and for what he needs to wear the clothes, and on whether he has an opportunity to return them to the shop. These examples indicate that even when the contract test is applied to actions governed by familiar moral principles, the correct conclusion may depend on relatively subtle differences in implications. Lacking information about such implications, or merely

underestimating or overestimating them, can lead one to form mistaken conclusions about the general acceptability of principles.

In the three types of cases considered, perspective-taking inaccuracy is not unlikely to result in mistaken conclusions about the justification of moral principles. Although empirical evidence does not show exactly how likely persons are to draw mistaken conclusions in these types of cases due to perspective-taking inaccuracy, the material discussed above provides sufficient reason to question its correct-usability with respect to such cases. This challenges the Practicability Assumption. Not only do these tricky cases appear to be quite common, one would hope that a moral guide can help in particular with regard to difficult questions.

The challenge is greater than just this. An agent may often, when considering a given action or principle, not be able to ascertain whether it an easy case in which the test is correct-usable or a tricky case in which it is not. In particular, when an agent either misunderstands or overlooks a standpoint, she will usually not be aware *that* she is misunderstanding or overlooking something. She may then reasonably think she is confronted with an easy case that clearly does or does not satisfy the contract test, even though she is in fact facing a tricky case in which she would have drawn the opposite conclusion had her understanding of the relevant standpoints been better.

The challenge is this. If agents cannot very well distinguish between easy cases in which the contract test is correct-usable and tricky cases in which it is not, they do know when they can trust judgments arrived at by the test and when they cannot. An agent who is aware that the contract test is not correct-usable with respect to tricky cases may therefore also not have reason to trust its conclusions in those easy cases in which it would in fact be correct-usable: for all she knows, it may actually be one of those tricky cases. Put differently, she may not have reason to rely on the contract test as a moral guide.

This challenge can be defused if there turn out to be ways in which persons may reduce the risk of drawing mistaken conclusions with respect to tricky cases. In that case, they may yet learn to use the contract test adequately, which is what the Practicability Assumption is committed to. This is the question of the next chapter.

6 Conclusions

Applying the contract test to evaluate a principle for the general regulation of behaviour requires one to identify alternative standpoints regarding the

principle than one's own present perspective and to determine what objections may be posed from them. This chapter examined what findings on our accuracy in perspective-taking reveal about our ability to do this adequately.

Against the concern that our perspective-taking ability is not at all up to the task, I argued that empirical findings suggest we are able to identify alternative points of view regarding moral principles and that we can achieve an understanding of these. This suggests we can apply the contract test successfully in particular cases.

That is, however, not yet to say that we can apply it adequately in general. Empirical findings show that one's ability to grasp an alternative perspective regarding an object is constrained by three factors: available cognitive resources, available information about differences between oneself and the occupant(s) of the target perspective, and one's own present perspective regarding the object. Given that both cognitive resources and information are typically limited, and that we are unlikely to fully inhibit the influence of our own present perspectives to our interpretations of other points of view, our perspective-taking accuracy is limited. I argued that this not only means that persons are prone to make certain mistakes when applying the contract test, but also that they may draw mistaken conclusions about whether principles would be the object of general agreement. In the previous section I identified three type of cases with respect to which this risk is particularly high. As such tricky cases are difficult to distinguish from easy cases in which perspective-taking errors are unlikely to yield mistaken conclusions, the existence of tricky cases provides a serious challenge to the contract test's practicability as a moral guide: agents may not have reason to trust conclusions they form with respect to easy cases because, for all they know, they may be facing a tricky case in which they cannot apply the test reliably.

Our perspective-taking ability is thus not such that we are naturally skilled at applying the contract test. This does not yet show the Practicability Assumption to be implausible, however. The Practicability Assumption requires not that agents can at present use the contract test adequately, but that they can, without too much difficulty, come to use it adequately. The next chapter discusses therefore what they can do to improve their performance with the contract test.

5

How to Use a Contract Test

1 Introduction

The question I set out to answer is whether agents can come to use the contract test adequately as a moral guide, as proposed by moral contract theorists such as Scanlon or Gauthier. This comes down to investigating a variant of the Practicability Assumption that states actual persons can, without too much difficulty, learn to apply the contract test adequately under circumstances that include those typical of everyday life. As the capacity to consider other points of view is crucially involved in applying the contract test, I have examined this assumption in the light of the findings on our capacity for perspective-taking. What do these findings reveal about the plausibility of the Practicability Assumption?

Let's start with the supporting findings. Chapter 3 concluded that persons do sometimes, in circumstances of everyday life, engage in reasoning that involves perspective-taking. As the contract test is a reasoning procedure that involves perspective-taking, this conclusion supports the assumption that persons can apply the contract test under such circumstances. Chapter 4 added that persons do sometimes accurately identify and understand alternative perspectives. This supports the idea, I argued in the previous chapter, that persons can apply the contract test adequately under circumstances of everyday life.

But there were also findings that sit less well with the Practicability Assumption. Grasping alternative perspectives requires attention, effort and information. Applying the contract test successfully in a particular case requires therefore that one has sufficient cognitive resources to consider other standpoints towards the relevant principle, and that one has sufficient information to identify objections that may be posed from them. There is good reason to think that in practice these requirements are often not satisfied.

Consider first the limitation posed by lack of cognitive resources. Chapter 3 concluded there are situations in which moral judgments are called for but persons do not have sufficient cognitive resources to engage in either reasoning or perspective-taking. Under such circumstances of low cognitive resources, persons cannot apply the contract test. In addition to this, Chapter 4 concluded there are circumstances under which persons may have sufficient cognitive resources to engage in reasoning that involves perspective-taking, but yet insufficient to apply it adequately. Applying the contract test successfully to a moral principle may often require considering multiple alternative standpoints, and reaching an adequate understanding of these requires a series of effortful adjustments to one's own point of view. Applying the contract test is to engage in a thought experiment, and a rather laborious one at that.

The number of situations in which agents can apply the contract test adequately is also restricted by the amount of information available to them. As I explained in the previous chapter, in order to identify standpoints from which principles that would allow a given action may not be acceptable, one needs information about others and their situations. While agents may typically have sufficient information about others who would be directly affected by their actions, theirs are not the only standpoints that should be taken into account when applying the contract test. The general acceptance of principles would also affect others they are not even familiar with. If agents lack relevant information about such persons and their situations, objections to a principle in question or even whole standpoints towards a principle may be overlooked.

A final restriction is posed by the fact that we must understand alternative standpoints always from our own present point view. Chapter 4 presented experiments showing that interpretations of alternative viewpoints tend to be egocentrically biased. When attempting to understand a target viewpoint regarding an object, persons are prone not to make sufficient adjustments to their own point of view and instead project aspects of their own view to the target. Besides insufficient effort, this phenomenon seems to be caused by limited information about differences between one's own characteristics and occupants of the target perspective, but also by difficulties in inhibiting aspects associated with one's own point of view.

What do these limitations mean for the Practicability Assumption? They suggest first and foremost that the contract test is not a decision procedure that we can apply for all our everyday choices. This need not be problematic.

It is not what moral contract theorists require the contract test to be nor what contract theorists such as Scanlon and Gauthier suppose it to be. As I explained in Chapters 2 and 3, the contract test does not need to fail as a moral guide if it cannot be applied whenever we need to form moral judgments. Agents should be able to apply the test sufficiently often, however, such that they can rationally adopt and internalise moral principles to guide them in situations in which they can't apply it.

Nevertheless, the above limitations also challenge that agents can use the contract test in this way. As we saw in the previous chapter, there are important types of cases in which perspective-taking inaccuracy is not unlikely to lead a person to draw mistaken conclusions when applying the contract test. This raises doubts regarding the contract test's correct-usability with respect to such tricky cases. If agents cannot apply the contract test reliably with regard to these cases, principles they derive with regard to them may not be valid. This affects their reasons for relying on the contract test with respect to tricky cases. But not only that. It is unlikely, I argued, that agents can reliably distinguish the easy cases in which they can apply the contract test adequately from the tricky cases in which they cannot. The existence of tricky cases therefore does not just affect their reasons for relying on the contract test with respect to those cases, but also their reasons for trusting the contract test as a moral guide in general.

It is not yet said, however, that agents cannot improve their performance with the contract test. Persons tend to make perspective-taking errors, and if these perspective-taking errors would occur when applying the contract test, invalid conclusions may follow. But there may be ways to reduce the probability of such errors occurring, or of them leading to mistaken conclusions. While it is unlikely that agents can fully overcome the limitations posed by their cognitive resources, information, and their own point of view, there may be ways in which they can cope with these limitations. That is what the present chapter considers.

It is worth noting that the limitations in our perspective-taking ability identified in the previous chapters do not just apply to the variant of the Practicability Assumption on which I concentrated. In particular, errors may also occur when the contract test is used for philosophical reflection. Of course, lack of cognitive resources poses less of a problem when it is used for such a purpose, as philosophical reflection can be carried out in what I called circumstances of high cognitive resources. Insufficient information about the

situations of others as well as constraints provided by one's own point of view can, however, just as well lead to unreliability here.[1]

I shall in the following sections discuss three methods by which agents may improve their performance with the contract test. I start with a rather obvious method directed at the limitation posed by lack of information, namely that of gathering additional information about other standpoints. I then turn to the possibility of drawing on the perspective-taking abilities of others. This method, I shall argue, may be particularly effective for reducing the likelihood of mistakes due to egocentric bias. The third method is that of internalising those moral principles of which one has concluded, by applying the contract test, that everyone has reason to agree to them. This method is particularly suited for coping with limited cognitive resources, but may also help with respect to the other limitations.

While these three methods may seem quite different, there are two guiding ideas in the following discussion. The first is that persons may improve their performance with the contract test by relying on third parties. The second is that they may improve it through proper preparation. The upshot of my investigation is that moral contract theorists should embrace these ideas, and require agents to adopt these methods.

It is worth emphasising that, in comparison with the previous chapters, the following discussion is more speculative and less based on empirical studies. While for each of the three methods there is some empirical support, more research needs to be done, in particular regarding their effectiveness.

2 Gather information about alternative standpoints

Lack of information is one of the main reasons why persons may overlook alternative standpoints towards a moral principle or fail to identify objections associated with them. Gathering information about others and their situations is thus an obvious way to decrease the probability of erring with the contract test. This section discusses several ways to improve our information about others. The aim is to assess to what extent persons can in this way reduce the risk of making mistakes with the contract test. I shall distinguish between improving one's understanding of alternative standpoints through gathering

[1] Note that this fits Hare's suggestion, with which I started this book, that his disagreement with Rawls about what persons in the original position would agree to stems in part from him and Rawls having different attitudes themselves.

information while applying the test and doing so by improving one's background knowledge.[2]

Say I am applying the contract test to decide whether I ought to help another person who asks for my assistance. Assume also that I lack certain relevant information about this person's standpoint, and that this would lead me to draw a mistaken conclusion. Can I prevent this by gathering additional information? There appear to be several things I can do. In so far as I have not already done so, I may through careful observation become better informed about aspects of the other person's situation that influence how he would be affected were I not to help him. Moreover, I may gather information through communicating with him. Communication with those who would be affected by one's actions can inform one about characteristics that would influence how they would be affected, including characteristics that may otherwise have remained out of view. For example, through conversation I may learn about the other's needs, and about the importance for him of being helped.

There is also an important role for third parties here. Third parties may have additional information about standpoints one ought to take into account. To take another example, say I am considering whether it is permitted not to donate to aid organisations such as Oxfam. Third parties, including media such as newspapers, may provide information about the situations of those depending on aid and how they would be affected were such aid to discontinue. This includes the information *that* these people exist. Indeed, consulting third parties may often be crucial for identifying points of view that one would overlook due to a lack of information.

Through observation and communication agents can thus extend the information available to them when applying the contract test. They can thereby improve their understanding of the principles appropriate to their situation, and of alternative standpoints towards these principles. Such

[2] At this point it is worth noting one type of information that does not appear to increase perspective-taking accuracy much: information that we are prone to make certain mistakes. Several studies have found that informing people about the danger of egocentric biases did little to increase their perspective-taking accuracy (Epley, 2008). For example, the previously mentioned (3§3.1) experiment on the false consensus effect found that neither informing people beforehand about this phenomenon nor giving them online feedback on the accuracy of their estimates while they were filling in the test improved their accuracy (Krueger & Clement, 1994). Another experiment in the same study showed that even if participants were informed that a large number of individuals had agreed to be willing to perform a certain task (i.e. wearing a sandwich board), they still tended to project their own decision to wear the board or not when predicting another's behaviour.

information gathering can thus reduce the risk of drawing a mistaken conclusion.

There are, however, obvious limitations to this method. For one, it may not be practically possible to gather information. There may be little or no opportunity for either observation or communication, for example due to constraints of time or distance. There may be no third parties available with relevant information.

There is also a more principled problem. Persons may often not be aware that they lack an important piece of information. A clear example of such an 'unknown unknown', to borrow a helpful phrase from Donald Rumsfeld, is when a person does not know that there is an alternative standpoint that he fails to take into account.[3] Or when he mistakenly thinks his interpretation of an alternative standpoint is accurate. If this is a person's predicament when applying the contract test to an action, he will see no reason to gather additional information.

Besides gathering information about alternative standpoints when applying the contract test, agents can also improve their background knowledge about the standpoints that may be occupied regarding moral principles. Such knowledge can then be relied upon when applying the contract test, reducing the need to gather information.

There are many ways in which agents can improve their background knowledge of the various standpoints that may be occupied with regard to moral principles. A straightforward way to learn about how a person's characteristics, including his situation, affect his standpoint is by actually putting oneself in various situations. If I have been in the situation of a person who needed help, I have a better idea of the implications of not being helped—and even more so if I did not receive the help that I needed. Simply by living our lives we learn can much about the different standpoints people can occupy with regard to an action. We have all on occasion needed help, been lied to, been called names, and even been harmed physically, and thus we have an idea of what it is like to be in such situations. When we are confronted with

[3] Rumsfeld introduced the phrase in a press statement he made in his role as United States Secretary of Defense, at February 12, 2002. It was part of the following analysis: "There are known knowns; there are things we know we know. We also know there are known unknowns; that is to say, we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know." See http://www.youtube.com/watch?v=GiPe1OiKQuk

another person in such a situation, we can draw on such information to gain a more accurate view of her perspective.

We can also place ourselves in situations that we do not tend to occupy. An interesting example in this regard, mentioned by Constanze Binder (2012), is a role-changing practice in which men and women engage in Oaxaca Mexico. On International Women's Day, men adopt the roles of their wives: they undertake walks to fetch water and firewood, prepare food, wash clothes, and care for the kids. Binder claims that this has made men much more understanding of the position of women and had favourable consequences for their struggle for women's rights.

While this appears to be a great way to increase one's understanding of alternative standpoints, there are evidently serious practical restrictions on it. A lack of time alone restricts how many situations one may put oneself in. Second, it may be too costly to put oneself in certain situations. For one, I can hardly be expected to gain experience with situations that are bad for me. For example, I can hardly be expected to place myself—for real, rather than in imagination—in the situation of a person living in extreme poverty. Third, and more fundamentally, my own characteristics constrain whose shoes I can put myself into. Despite these restrictions, Binder's example does, however, nicely illustrate that we can do much more than we tend to do.

As with gathering information when applying the test, agents can improve their background knowledge of generic standpoints towards actions and principles through observation and communication. Third parties again play a crucial role. We can, and in fact do, share our observations about the implications of actions on persons in varying situations. This includes non-fictional media. Newspapers, magazines and television programmes, in particular those with a 'human interest', can teach us about situations which we are unlikely to observe with our own eyes, even though they may be relevant to take into account when thinking about our actions. Take for example news stories that inform us that certain products in our stores have been produced through a process that involves exploitation. But we may also learn from fiction. Many fictional stories deal with how the actions of one person affect others. The situations of fictional persons may resemble those of actual persons, and can thus provide information about standpoints. For example, films such as 'Guess Who's Coming to Dinner' (1967) and 'Brokeback Mountain' (2005) gave viewers more insight into the difficulties faced by romantic couples in an intolerant climate, and may as such reduce intolerance towards interracial and homosexual romantic relationships.

There are several advantages that increasing one's background information about standpoints has over gathering information about them when applying the test. First, it does not have equally strong practical restrictions. Agents have a lot of time, in the course of their lives, to acquire information about others and their situations. Second, acquiring background information reduces the probability of there being unknown unknowns when applying the contract test. Background information about others and their situations can direct agents towards standpoints of which they would otherwise have been unaware.

The considerations put forward in this section show that agents can improve their understanding of alternative standpoints through gathering information about them. That is not so say, however, that the limits posed by lack of information can be fully taken away. First, relevant information will sometimes be inaccessible. There may be no opportunity to communicate with those who would be affected by one's actions, for example. While preparing oneself by gathering relevant information beforehand can reduce the chance of this occurring, one can hardly prepare oneself for every situation.

The second reason is, once again, provided by our limited cognitive resources. Gathering information is effortful and time-consuming. As agents cannot prepare themselves for every possible situation, it remains the case that they sometimes need to make decisions without all the relevant data. Furthermore, even for cases in which agents can spend large amounts of time and cognitive resources in gathering all the relevant information, this may be expecting too much of them. While it can be argued that, given the importance of accurate moral judgment, agents should put a substantial amount of effort and time into getting all the relevant information, there are other valuable goals they may need to attend to.

3 Use the perspective-taking abilities of third parties

The previous section mentioned that agents can improve their knowledge about other standpoints by acquiring information from others. But there is a more fundamental way in which they can improve their contract test performance by drawing on the minds of others: they can make use of other persons' perspective-taking abilities. I shall start with explaining how this would work, and why it would work: why it can reduce the probability of errors occurring when applying the contract test.

Say I am, once again, using the contract test to decide whether I ought to help another person who asks for my assistance. Assume also that I would, if I were to depend solely on my own perspective-taking abilities, end up with an egocentrically biased interpretation of another standpoint. For example, I would underestimate the implications of not being helped because, had I been in the other's situation, there would have been an alternative course of action available to me that is unavailable to persons occupying the standpoint in question. Asking a third party to consider the other's standpoint can prevent me from making this mistake. If the aforementioned social psychologists are correct (4§3.2), I would be making this mistake by failing to take into account relevant differences between myself and the standpoint under consideration. The third party may not make that same error. First, she may have information about the standpoint that I am unaware of. Second, as she is differently situated than I am, she may not be prone to the egocentric biases that cloud my judgment.[4] She may be more similar to the persons occupying the target standpoint, and therefore have a more accurate understanding of the implications that the person in question would endure were I not to help him.

Relying on the perspective-taking abilities of others is not confined to concrete moral situations or individual standpoints. Agents can ask others for their evaluation of moral principles. Do others persons consider a given principle for the general regulation of behaviour one that everyone has reason to accept? Moreover, agents can evaluate their moral principles in collaboration with others rather than individually. Persons can go through the steps of the contract test together, correcting and completing each other on the way. To the extent that they have different characteristics and different information, this may seriously reduce the risk that alternative standpoints towards the principle or objections associated with such standpoints remain unaccounted for.

Using the perspective-taking abilities of third parties has some advantages over merely asking them for information. The first is that it is more likely to correct egocentric biases in one's interpretation of an alternative standpoint. When a third party informs me of the needs of a person whose perspective I am considering, I shall integrate this in my interpretation of that person's standpoint. An existing egocentric bias is unlikely to be fully removed by such information (Epley, 2008). When on the other hand the third

[4] There is also an issue of motivated reasoning. There is good reason to think that our judgments tend to be influenced by our self-interest. Third parties who do not share our self-interest can correct such a bias.

party would form a judgment about the alternative standpoint with regard to a principle under consideration herself and inform me of this judgment, an egocentric bias in my own interpretation may be made explicit, presuming this third party does not make the same mistake.

The second advantage is that relevant information available to that third party is more likely to be taken into account. When I use the third party as a source of information, I may not ask all the relevant questions and thus not gather everything there is to know. Were I to let her make her judgment herself, such information is more likely to be included. This is particularly helpful when the third party is much better informed than I am. A third, but less prominent advantage, is that it can be relied upon when one does not have sufficient cognitive resources to apply the contract test oneself but the other does. In such cases I can choose to rely on the judgment of the other.

While using the perspective-taking abilities of third parties in general reduces the likelihood of making mistakes when applying the contract test, it is far from a panacea. For one, it is evidently not a method that agents can use whenever they apply the test. There may be no opportunity to ask third parties, either due to constraints of time or because there are no (willing) third parties available. Furthermore, third parties may just as well be victim to egocentric biases resulting from their specific positions on this issue. While this should not pose a risk of error in case agents are not prone to these biases themselves, either because they are sufficiently informed or because they do not share the characteristics that underlie these biases, it may do so when they are not better situated than the third-parties in question. Indeed, in that case the judgements of third-parties may even *strengthen* mistaken judgments. I take this to mean that agents must be careful which third parties they choose to ask for advice. It may be good practice to choose reasonable others who are differently situated from oneself, as these are less likely to have the same egocentric biases.

It may be thought that the idea of drawing on the perspective-taking abilities of others is in conflict with the individualist nature of the social contract theory. Both Scanlon and Gauthier take judgments about which principles are justified to be judgments that we make as individuals. As is well-known, this view has been criticised by Jurgen Habermas. On Habermas's view, justification of moral principles needs to be sought through social reasoning in real discourse rather than through an individual's reasoning about hypothetical agreement. Scanlon has explicitly rejected Habermas's view. As Scanlon writes,

> while interaction with others plays a crucial role in arriving at well-founded moral opinions [...] reaching a conclusion about right and wrong requires making a judgment about what others could or could not reasonably reject. This is a judgment that each of us must make for him or herself. The agreement of others, reached through actual discourse, is not required, and when it occurs does not settle the matter. (Scanlon, 1998, pp. 393-394)

To see that my proposal is not in conflict with the position that Scanlon puts forward here, it is important to emphasise that under my proposal one draws on the perspective-taking abilities of others *not* with the goal of agreeing with them but with the goal of improving one's own judgment. One remains the final arbiter oneself. My proposal is therefore fully compatible with this individualist aspect of contract theory. To the contrary, while Scanlon does not say much more about it, at least not in *What We Owe to Each Other*, my proposal fits nicely with his view that "interaction with others plays a crucial role in arriving at well-founded moral opinions" (1998, p. 393).

I conclude that using the perspective-taking abilities of third parties may be an effective method to reduce the risk of error when applying the contract test. By drawing on the minds of others, agents can overcome limitations and biases associated with their own imagination.

4 Internalise principles that would be the object of agreement

Without sufficient cognitive resources, persons cannot apply the contract test adequately. In Chapter 3 I argued that this need not imply that it cannot be relied upon as a moral guide for those situations. Persons may be able to internalise moral principles that satisfy the test and follow these when their cognitive resources are insufficient. I argued there also that empirical findings on the diachronic role of perspective-taking in moral judgment suggest that persons can indeed shape their future moral judgments by applying the contract test.

In the present section I shall discuss the plausibility of such a response further. I start by explaining how other moral theorists, and utilitarians in particular, have responded to the charge that persons cannot use their standard of permissibility to evaluate actions. I will argue that contract theorists can adopt a similar response, and more convincingly so. I will then argue that internalising moral principles that satisfy the contract test may not only enable persons to cope with their limited cognitive resources, but may also be expected to reduce the need for information gathering and the occurrence of errors of egocentricity.

The problem of limited cognitive resources is also faced by act utilitarians. As Mill (1871/2001) writes in *Utilitarianism*, "defenders of utility often find themselves called upon to reply to such objections as this—that there is no time, previous to action, for calculating and weighing the effects of any line of conduct on the general happiness" (p. 23). In addition, utilitarians have recognised that we are likely to make mistakes when calculating the utility of courses of action. Given that utilitarians face a similar problem, their responses to it may be helpful for contract theorists.

From Bentham onwards, utilitarians have advised us to choose our actions not on the basis of the principle of utility but instead to rely on rules that tend to lead to the best possible state of affairs. Two-level utilitarianism as developed by Hare is an extreme variant of this. On Hare's view, using some version of the principle of utility as a guide for everyday behaviour requires "superhuman powers of thought, superhuman knowledge, and no human weaknesses" and is as such only possible for "archangels". We should therefore rely on "intuitive moral thinking" rather than "critical moral thinking" in everyday situations (Hare, 1978). In such situations, we should follow those moral rules and practices that we would endorse when engaging in critical moral thinking in a cool hour. Only when we face a difficult moral scenario, such as one in which moral rules appear to conflict, should we form our judgment through critical moral thinking.[5]

Contract theorists can put forward a similar proposal. As I mentioned before, while moral contract theorists such as Gauthier and Scanlon require persons to be able to act in conformity with principles that satisfy the contract test, they do not require them to choose all their individual actions on the basis of it. Indeed, given that they justify actions in terms of principles this move fits much more naturally with contract theory than it does with act utilitarianism.[6] By following principles that satisfy the contract test, persons would do exactly what the contract test prescribes.

Not everyone has been convinced by the utilitarian's response to the problem of limited cognitive resources. It is worth noting, however, that certain important arguments voiced against two-level utilitarianism do not apply to contract theory. A first objection posed against the utilitarian's response is that two-level utilitarianism undermines an agent's commitment to

[5] It is worth noting that this proposal is not the same as rule utilitarianism. Rule utilitarians hold that an action is justified if it conforms to rules that maximise utility. As making such a judgment requires extensive calculations as well, it is not a solution to the above problem.
[6] Rule utilitarianism does have this same advantage over act utilitarianism.

act in accordance with her moral standard. A utilitarian moral agent would know that the rules that he follows in everyday situations are a 'mere' guideline, an approximation of the real standard. A related objection is that the two-level view requires an agent to engage in two conflicting ways of thinking: to switch between a non-consequentialist mode of thinking and a utilitarian one (both objections are discussed in McNaughton, 1988, pp. 180-181).

It is not hard to see that these objections do not apply to the contract theorist's version of the response. The contract theorist has persons following principles that satisfy the contract test rather than approximate it. Whether a contractarian moral agent follows such principles or evaluates actions by applying the contract, actions are justified in the same way: by being in conformity with principles that would be the object of agreement.[7]

However, there is another objection to the utilitarian's response that applies just as well to the contract theorist's version. By letting agents in certain situations follow principles rather than apply a procedure of justification in order to choose their actions we assume that they can, at a certain point in time, learn these principles. Applying the contract test to all possible instances requires a lot of effort and time, and remembering its conclusions appears to require a very good memory. There are many different situations and many different moral principles that govern them. Indeed, as I mentioned before (2§3), Scanlon suggests there is an "indefinite number" of valid moral principles. It is therefore not possible for agents to prepare themselves for all or even most situations in which they cannot apply the contract test, the objection would be.

Contract theorists have two responses to this problem. One response, mentioned in the introduction of this section, is based on findings presented in the previous two chapters. Agents can, in their everyday lives, apply the contract test to evaluate moral principles. By internalising such conclusions they can develop a moral understanding on which they rely in the future. But before discussing this idea further, I turn to another response which is again inspired by how utilitarians have responded to a related objection.

---

[7] Gary Watson (1998)has also argued that contract theories do not face the same problems as two-level views, but he concentrates on a different issue. To alleviate the objections posed here, some utilitarians have suggested that it may be better if persons were taught moral rules without becoming whole-hearted utilitarians. "Just people must be [...] to a certain extent deluded about the grounds of their virtue" (p. 173) Watson concludes that two-level views in this way exclude one of our deepest moral aspirations, something which contract theories do not need to do.

In response to the objection that persons cannot on every occasion determine the consequences of their actions, Mill writes:

> The answer to the objection is, that there has been ample time, namely, the whole past duration of the human species. During all that time, mankind have been learning by experience the tendencies of actions; on which experience all the prudence, as well as all the morality of life, are dependent. (Mill, 1871/2001, p. 23)

Mill points out that determining the consequences of our actions on happiness is not something for which we have to start with from scratch. The generations before us have acquired extensive knowledge regarding the consequences that actions tend to have for happiness. This knowledge, Mill writes, has been laid down in the moral precepts that make up "the morality for the multitude". Mill goes on to argue that, as long as we have no good reason to think that such moral precepts do not conduce to our general happiness, we should follow them.

This response includes two elements of relevance for the plausibility of the contract theorist's proposal. The first is the observation that almost all of us are able to internalise society's moral precepts, and follow these rules in their everyday decision-making. This provides reason for thinking that agents can also learn to follow the complex system of rules justified by the contract test. This may be a long-term learning process that requires time and effort, from both the agent and parties that play an educating role, but it is something that most of us manage to do.

The second element is that through socialisation in society, persons have already internalised many of the principles that they ought to follow. They therefore do not have to explicitly derive all valid moral principles in order to prepare themselves for moral situations. Both Mill and Sidgwick suggest that the morality of common sense, as Sidgwick calls it, is close to the set of rules that utilitarians should follow in their day-to-day dealings. Contract theorists can make roughly the same argument. What is more, the argument is in their case more plausible.

A well-known objection to act utilitarianism is that it does not fit our moral intuitions. In particular, it does not fit our intuitions regarding deontic constraints (Watson, 1998). Most of us hold that persons have rights which constrain how we may pursue the good. We thereby hold that there are situations in which it is not allowed to maximize aggregate well-being. For example, we hold that persons have a right not to killed, a right not to be deceived, and a right to privacy, and that these rights must be respected even when not doing so would maximize aggregate well-being. We hold that when

a person makes a promise to another she grants him a right, which she may not break solely because doing so has better consequences. These intuitions cannot be accounted for by act utilitarianism. It holds that the rightness of actions depends solely on their consequences for the total well-being, and thus denies the existence of deontic constraints. As deontic constraints have a crucial place in common sense morality, this places some doubt on the claim that one can be a utilitarian by complying with society's moral precepts.

Contract theories do not face this problem (Watson, 1998). Contract theorists hold that persons must follow principles that would be the object of hypothetical agreement. While the well-being of individuals plays an important role in the justification of principles, these principles constrain how persons may pursue the good. As Watson puts it, "the agreement will include rights-conferring principles restricting how we may treat others without their consent" (p. 256). The agreement may be expected to include principles that forbid persons to kill, to deceive, and to invade the privacy of others, thus conferring on them rights in these matters. In contrast to utilitarianism, contract theories affirm the existence of deontic constraints. Contractarian morality is therefore in this important respect closer to the morality of common sense than utilitarian morality.[8]

Nevertheless, it is not to be expected that common sense morality and a contractarian moral conception are exactly identical. This is where the other response to the objection that agents cannot learn all valid principles comes in. Whereas it is unlikely that agents can derive all valid moral principles by applying the contract test in the occasional cool hour, it is not implausible that they can learn such moral principles through actual moral practice. I have argued that findings on perspective-taking suggest that they can apply the contract test successfully in many everyday situations, individually or in

[8] It may be objected at this point that contractarian morality is in other respects less similar to common sense morality than utilitarian morality. By deriving moral principles from an agreement by rational parties, contractarians may appear unable to take into account the interests of individuals who are not able to take part in such an agreement, such as the unborn, the congenitally handicapped, or animals. Many of us, however believe that these groups also have certain rights. Gauthier has stated explicitly that his moral theory can indeed not do justice to such intuitions. On his view, the above groups "fall beyond to pale of a morality tied to mutuality" (p. 268). Other contract theorists have argued that this conclusion is, even for Gauthier, by no means necessary (Hampton, 1991; Morris, 1991). As Morris (1991) explains, that individuals cannot partake in an agreement does not preclude them from being granted rights through principles agreed upon by rational others. Furthermore, as Scanlon (1982; 1998) shows, contract theorists who do not aim to derive morality from rationality can allow for individuals of these groups being represented by trustees in the agreement situation. The objection thus fails.

collaboration with others. If agents have many opportunities to evaluate moral principles, they also have many opportunities to prepare themselves for those situations in which they cannot apply the test due to a lack of cognitive resources.

One aspect of the above worry is that agents are unable to remember a large number of moral rules. In response to this, it must be emphasised that the result of preparation or moral education does not need to come in the form of *explicit* rules. Consider again what Scanlon writes about what it is to understand moral principles, in which he concentrates on the principle that promises freely made must be kept:

> Anyone who understands the point of promising—what it is supposed to ensure and what it is to protect us against—will see that certain reasons for going back on a promise could not be allowed without rendering promises pointless, while other exceptions must be allowed if the practice is not to be unbearably costly. [...] All of this structure and more is part of what each of us knows if we understand the principle that promises ought to be kept. In making particular judgments of right and wrong we are drawing on this complex understanding, rather than applying a statable rule, and this understanding enables us to arrive at conclusions about new and difficult cases, which no rule would cover. (Scanlon, 1998, pp. 200-201)

To understand a moral principle, I take Scanlon to be saying, is not to know a statable rule about a certain type of action. Understanding a moral principle is to have a complex structure of knowledge of standpoints and reasons associated with a type of action. When confronted with actions to which the principle applies, this knowledge, which may be largely implicit, points agents to the relevant considerations and enables them to quickly form moral judgments. There is no need to explicitly apply the contract test.

This moral understanding may be associated with having certain moral intuitions. As I described before (3§3.1), reasoning can affect and shape the moral intuitions that we have at a later stage. The diachronic role of reasoning that involves perspective-taking in moral judgment supports the idea that the contract test can also be used for this purpose (3§4). By applying the contract test under favourable conditions and drawing conclusions about moral principles and moral ideas, agents may thus not only develop a moral understanding but also develop associated automatic responses that can guide them in situations in which there is no opportunity to reason.

Together, these two responses answer the objection that agents may not be able to learn the system of moral principles justified by the contract test. They may be expected to already know many of the principles that satisfy the contract test, including standpoints and reasons associated with them. By

applying and learning from the contract test in cognitively favourable contexts, they can shape and develop this moral understanding further.

Internalising principles that satisfy the contract test appears to be an effective way to act in conformity with such principles in situations in which one's cognitive resources are too low to apply the test. But that is not the only beneficial effect of internalising principles. By internalising moral principles, agents may reduce the need for information gathering in future cases. Scanlon hints at this in the above quote when he says that understanding of moral principles enables us to arrive at conclusions about new and difficult cases. An example of such a case is when two familiar moral principles conflict, such as when one must choose between voicing a harmless lie or hurting another with the truth. The right choice in such cases depends on the details of the situation. But knowledge of associated moral principles certainly helps in deciding what to do. It presupposes an understanding of the relevant standpoints and considerations associated with these standpoints. The need for considering alternative standpoints is thus reduced, and with that the need for gathering information and spending cognitive resources. Furthermore, this moral understanding should direct one's attention towards those standpoints that may have most reason to object to the principles under consideration, and thus facilitates the deliberative process as well as the search for relevant information pertaining to the specific case.

Internalisation may also reduce the occurrence of errors resulting from the inherent egocentricity of our perspective-taking ability. Say I consider whether I may break a promise to another person, the keeping of which is somewhat disadvantageous for me. In such a case, it is not unlikely that my own inclinations towards the action would colour my interpretation of the objections that may be posed from the other person's standpoint. This effect may be so strong that I would mistakenly conclude that the breaking of my promise is permissible. An internalised moral principle may have a correcting influence in such cases. I would in that case already know that the other has a strong objection against my breaking my promise before an egocentric bias could make me think otherwise.

To conclude, there is reason to think that even though agents cannot apply the contract test *in* certain situations, the contract test can be a moral guide *for* those situations. By deriving conclusions from the contract test under conditions more favourable to our perspective-taking ability, agents can develop a moral understanding and associated intuitions on which they can rely with respect to similar situations. Internalising principles that satisfy the

contract test appears to be an effective method to cope with our limited cognitive resources. In addition to this, as an understanding of moral principles implies an understanding of relevant standpoints and associated objections, it may also be expected to improve one's ability to apply the contract test when confronted with new and difficult cases related to these principles.

5 Is the Practicability Assumption empirically plausible?

This chapter started with the observation, derived from the previous two chapters, that empirical findings do not show that actual persons can use the contract test adequately as a moral guide. To the contrary, empirical studies reveal their perspective-taking ability to be limited in ways that may affect agents' performance with the contract test negatively. The question this chapter set out to answer is whether they are able to do something about this. The Practicability Assumption namely requires not that actual persons can use the contract test adequately, but that they can *come* to use it adequately. The previous three sections discussed three general ways in which agents may cope with the limitations of their perspective-taking ability. In this final section I will discuss what this means for the plausibility of the Practicability Assumption.

Let me start with what may be the largest limitation, that agents sometimes have insufficient cognitive resources to apply the contract test. I discussed two methods that can help them cope with this limitation. The first is to rely on the perspective-taking abilities of others: *they* may be able to apply the contract test. Of course, this only works when trustworthy others with sufficient cognitive resources are available. The second and more widely usable method is to prepare for such situations by internalising moral principles under conditions more favourable to the ability for perspective-taking. There is reason to think that through socialisation and moral education persons have already internalised an important part of these principles. They can in that case use the contract test to shape and further develop this existing moral understanding.

This method should also be of help for a closely related limitation, namely that agents' cognitive resources are sometimes so low that they would make mistakes were they to apply the contract test. By internalising a given moral principle agents may decrease the cognitive resources required for applying the contract test with regard to situations to which that principle applies: as

knowing the moral principle presupposes an understanding of the relevant standpoints and objections associated with the principle, processing costs should be reduced. While this information may not be sufficient to yield conclusions for new and difficult cases, understanding of principles related to such a case should be helpful. Such understanding means there is less 'new' to consider, and may direct their attention to the crucial considerations.

Agents may thus to a great extent be able to cope with the limitation caused by cognitive resources. What about the limitation posed by lack of information? As moral principles include information about standpoints and objections, internalised moral principles should reduce the need for information gathering, I argued in the previous section. However, rationally adopting the principles in the first place does of course require agents to have access to this information. Without sufficient information about others and their characteristics, they will not reach an adequate interpretation of alternative standpoints towards principles. As I explained in the previous chapter (§4-5), this may lead them to draw mistaken conclusions when applying the contract test.

The most straightforward way to deal with this limitation is to gather additional information. In section 2 I distinguished between gathering information about alternative standpoints while applying the contract test, and developing background knowledge regarding standpoints towards principles. With regard to the former, I argued that there are several ways through which persons can become better informed than they tend to be. Communication with others, including those who would be differently affected by principles under consideration, may be the most important of these.

Agents may, however, be unaware of their ignorance of alternative standpoints and associated objections. As they would in such a case see no reason to gather additional information, this method is of limited use. I argued that for this reason, as well as for practical concerns, agents should also develop their background knowledge of standpoints associated with the principles. I mentioned several ways in which they can do this, including actually placing themselves in other situations. Less demanding ways include carefully observing others in different situations, communicating with them, following the news and documentaries, and studying the situations and standpoints of fictional characters. By becoming better prepared for moral situations, there are less likely to be unknown unknowns.

Another method that may decrease lack of information is that of drawing on the perspective-taking abilities of third parties. Instead of applying the

contract test by oneself, agents may call in the help of others, or even apply it in collaboration. An advantage of this method is that others may have information about alternative standpoints unavailable to agents themselves, including information they would not consider asking about.

Using the perspective-taking ability of others may also work against the third cause for inaccuracy, that interpretations of other points of view tend to be coloured by agents' own present perspectives. Due to having different characteristics or even being differently situated with regard to the principle in question, third parties are unlikely to be affected by identical egocentric biases. Appealing to another's perspective-taking ability may thus make agents attentive to egocentric aspects of their own interpretations. Note that egocentricity may also be reduced by the other methods. Egocentricity results because persons make insufficient adjustments to their own point of view when interpreting those of others. Having more information about differences between themselves and others increases persons' awareness of which adjustments must be made, whereas increases in cognitive resources improve their ability to make these adjustments. Finally, internalised valid principles, which include information about standpoints and associated objections, may correct egocentric interpretations.

There is thus quite a lot that agents can do to improve their ability to apply the contract test. By extending their knowledge about other standpoints through information gathering and the internalisation of moral principles, they can improve their performance with the contract test. Furthermore, they can discuss difficult and new cases with others, combining their perspective-taking powers and apply the contract test collaboratively. This should seriously reduce the probability of drawing mistaken conclusions. In particular, I take it that the combined use of these methods should enable agents to also arrive at correct conclusions with respect to the tricky cases that I identified in the previous chapter (§5). And as agents can internalise such conclusions, they do not have to do each time again.

That is not to say that we have reason to think agents can become perfect users of the contract test. Perspective-taking inaccuracies are likely to remain with us to a certain extent, and agents will as a consequence sometimes draw mistaken conclusions about whether principles would be the object of agreement or not. Furthermore, there are likely to remain difficult or new cases that we cannot solve with our moral understanding, but in which we also do not have sufficient resources to apply the contract test successfully. While these are limitations on the contract test's practicability, they do not imply it is

not an adequate instrument for moral justification. As I said when discussing criteria for such instruments (2§2), it seems too much to ask of them to provide a determinate solution in every situation and be correct-usable to the point that a designated user never errs when applying it.

So what should be the verdict on the Practicability Assumption's empirical plausibility? Although there will remain a risk for making mistakes when applying the contract test, the above considerations provide reason to be optimistic about agents' ability to cope with the limitations of their perspective-taking ability and decrease this risk substantially. I thus conclude that, at least with respect to findings on social cognition, the Practicability Assumption is empirically plausible.

This is good news for contract theory. It supports the idea that it lies within our power to act in conformity with principles that would be the object of agreement. Persons can, provided they adopt the methods presented in this chapter, become contractarian moral agents. Given this provision, contract theorists should, I believe, emphasise the importance of adopting these methods.

II

The Translucency Assumption

Do I know the difference between right and wrong, and do I want to be good? Sure. One catches more flies with honey than with vinegar. A peaceful and orderly world is a more comfortable world for me to live in. So do I avoid breaking the law because it's 'right'? No, I avoid breaking the law because it makes sense. I suppose if I weren't gifted with the ability to make a lot of money in a profession doing what I like, I might try and profit by crime. But with my profession, I'd have to really hit the criminal jackpot to make it worth a life of crime.

Anonymous psychopath, 2010

6

Contract Theory and Translucency

1 Introduction

One of the crucial questions that a normative theory of morality must address is why persons would be moved to act on its demands (Freeman, 1991; Scanlon, 1982). As Freeman (1991) puts it, "for any philosophical account of morality to be convincing, it has to connect awareness of moral requirement with action" (p. 289). It must explain why persons who understand the moral requirements it puts forward would be motivated to act on them. As I have already explained in the introductory chapter, contract theorists of the Hobbesian strain answer this question differently from contract theorists who take after Kant. I shall now briefly explain the main difference between their answers, which will serve as an introduction of the Hobbesian contract theorist's approach on which I will focus. The remainder of this chapter will explain David Gauthier's particular answer.

Kantian contract theorists such as Rawls and Scanlon assume persons are in fact to a certain degree motivated to comply with moral demands. Rawls (1999) takes this to be an aspect of the sense of justice that he ascribes to reasonable persons, "an effective desire to comply with the existing rules and give one another that to which they are entitled" (p. 274). Similarly, Scanlon writes that "the source of motivation that is directly triggered by the belief that an action is wrong is the desire to be able to justify one's actions to others on grounds they could not reasonably reject" (p. 116)[1]. Both authors argue extensively that persons with these desires will normally be motivated to comply with the demands of their respective moral conceptions. Importantly, for Kantian contract theorists this is not just any desire, but one that persons

[1] In his later work *What We Owe to Each Other* Scanlon still holds that persons are motivated to justify themselves, but no longer holds that this motivation should be understood as a desire. This change does not affect the point made here.

have very good reason to have and that moderates and limits their other desires. On Scanlon's view, for one, the desire to justify oneself to others is based on the intrinsic value of standing in a relation of mutual respect with others, which he takes to explain the priority of moral considerations over other reasons.

Hobbesian contract theorists approach the motivational question rather differently. As I explained in Chapter 1, Hobbesian contract theorists aim to ground morality in instrumental rationality. Clearly, such an account cannot not rely on pre-existing moral desires that not every rational being needs to have. Hobbesian contract theorists, and Gauthier most famously, therefore seek to show that morality is an effective way to pursue one's amoral aims and interests. This would allow for an obvious connection between awareness of moral requirement and action: in so far as persons may be expected to be moved by self-interest, they may be expected to be morally motivated.

The crucial difference between these two approaches is perhaps seen most clearly in what they have to offer to an agent who does not already care about morality, an amoral agent. A Kantian contract theorist such as Scanlon has no ambition of convincing the amoral agent. Indeed, when mentioning the possibility of "justifying the morality of right and wrong to someone who does not care about it—an 'amoralist'", Scanlon states that, "I myself doubt whether such a justification can always be provided" (p. 148). He takes it for granted that moral concerns move the great majority, and that as such it is sufficient to give "a fuller explanation of the reasons for action that moral conclusions provide" (p. 148). A Hobbesian contract theorist such as Gauthier, on the other hand, takes the amoral agent as his prime target. He aims to show that amoral agents have reason to adopt principles that everyone has reason to agree to. In that case, he would have shown persons have reason to be moral given solely their amoral aims and interests.

An advantage of Gauthier's approach to the motivational question is that self-interest is undoubtedly a major source of motivation. This is not evidently the case for moral concerns such as the desire to be able to justify oneself to others, which may not be universal and may not be very strong. As self-interest does appear to be a strong source of motivation, one may question whether it would not often override such moral concerns. In turn, one may worry that it would be difficult for agents to comply with moral conceptions that identify such moral concerns as their "motivational basis", such as is the case for Scanlon's (1982; 1998) contractualist conception. In that case, the conception may also not be able to play an effective social function; it may not

have the requisite stability (cf. 2§3). If successful, Gauthier's project may alleviate such, as it would show that the conflict between self-interest and morality is less deep than it is usually thought. As he puts it, "duty overrides advantage, but the acceptance of duty is truly advantageous" (p. 2).

The following chapters will consider the empirical plausibility of Gauthier's claim that it is advantageous to be moral. I shall concentrate in particular on Gauthier's empirical assumption that people are to a certain degree translucent: that we can detect one another's moral disposition. In this first chapter I will explicate the assumption and address how it can be investigated. As Gauthier makes the assumption that people are translucent in the course of answering the so-called Compliance Problem, I shall start with discussing this problem. I then turn to assessing the exact role of the assumption in Gauthier's argument. I finish with discussing some important objections to the assumption, which I will use to structure the investigation in the following chapters.

2 The Compliance Problem

In *Leviathan*, Hobbes argues that given how bad it is to be in a state of nature in which everyone does as they like, each of us would be better off if everyone accepted certain moral constraints on the direct pursuit of their interests. Therefore, Hobbes argues, each of us has reason to adopt these constraints. This argument lies at the basis of the Hobbesian strain of contemporary contract theory of which David Gauthier is the best-known defender.

Like Hobbes, Gauthier holds that we would be worse off in a world without morality. He explains this by connecting morality with cooperation. More than other animals, humans can benefit from working together. Take for example cooperative hunting. We are much more likely to catch a large animal if we cooperate than if we act individually. It thus makes sense for hunters to agree to work together and share the benefits of this cooperative activity. However, once one of us has caught the prey due to the help of others, it may not be in his interest to share it with the other parties. Cooperative activities often involve situations in which it is in one's interest to depart from the mutually advantageous cooperative activity. They often involve an incentive to engage in *free-riding*, to exploit the cooperative behaviour of others. I call such situations *self-benefiting opportunities*. If parties who may engage in a cooperative activity are aware that others will have self-benefiting opportunities, they may not even recognise an incentive to cooperate in the

first place. This is why there is use for morality, Gauthier claims. The mutual acceptance of norms that forbid taking advantage of others enables us to engage in cooperative activities with one another. On Gauthier's view, "moral principles may be understood as representing [cooperative action] prescribed to each person as part of the ongoing co-operative arrangements that constitute society" (p. 168).

Gauthier thus conceives of morality as a set of social norms that govern cooperative practices. That is not to say that every social norm counts as a moral norm. First, besides norms of cooperation, which include moral norms, there are norms of conformity (Tomasello, 2009). It is typical of norms of conformity that their violation does not involve the exploitation of others, even though it may disturb the social practice which it governs.

Second, and more importantly for our purposes, norms of cooperation may not count as moral for failing to satisfy certain conditions that are required for them to be *justified*. As I explained in Chapter 1, Gauthier identifies morality as those norms or principles that idealised representatives of ourselves who would be bargaining about their terms of interaction under certain idealised conditions would come to agree on. In order to ensure that the principles chosen are mutually advantageous, Gauthier assumes these representatives to be rational, in the sense that they maximise their expected utility, and to have no moral preferences. In addition, to avoid the content of the principles depending on sentiments not shared by everyone, Gauthier assumes that these representatives are mutually disinterested: they take no interest in the interests of others. This also ensures that the principles do not favour certain individuals due to the affections others have for them, or exploit certain individuals due to the affections they have for others (Hampton, 1993). The crucial idea is that if such persons would agree to a given set of principles, this set of principles is also rational for us to agree to given our non-moral aims and interests. Besides these conditions, Gauthier adds an additional condition to ensure that agreed upon norms are not only mutually advantageous but also impartial (the 'Lockean Proviso'). In this and the following chapter, I will take moral norms to be those cooperative norms that satisfy Gauthier's contract test.

But as Gauthier points out, in order to show that, given our amoral aims and interests, we have reason to accept such principles as constraints on our behaviour, it must not only be the case that such representatives would agree to these principles but also that they would stick to the agreement made in their interactions with one another. This requires an additional argument, as it

does not follow from the fact that an agreement is advantageous to make, that it is also advantageous to keep to it. It may be advantageous to make an agreement at one point in time but not advantageous to keep to it at a later point in time. As such, it may also not be rational to keep it.

Hobbes recognised this problem, and introduces it by means of his antagonist 'the Foole':

> The Foole hath sayd in his heart, there is no such thing as Justice; and sometimes also with his tongue; seriously alleaging, that every mans conservation, and contentment, being committed to his own care, there could be no reason, why every man might not do what he thought conduced thereunto: and therefore also to make, or not to make; keep, or not keep Covenants, was not against Reason, when it conduced to ones benefit. (Hobbes, 1651/1991, p. 101)

The Foole's logic appears impeccable. If what is rational depends on what is in one's interest, and it is in one's interest to violate an agreement that was in one's interest to make, then it is "not against Reason" to violate the agreement.

In order to show that people have reason to be moral given their amoral aims and interests, Gauthier needs to refute the Foole. It needs to be shown that people have reason to comply with the principles they have reason to agree to. This problem is often called the Compliance Problem.

The Foole's moral scepticism does not just apply to Hobbesian contract theories such as Gauthier's. The Foole denies in general that there is reason to be moral. However, Hobbesian contract theorists such as Gauthier must take the Foole particularly seriously. Like the Foole, Gauthier holds that morality must be justified in order to appeal to amoral aims and interests. As he puts it himself: "The Foole challenges the heart of the connection between reason and morals that both Hobbes and we seek to establish—the rationality of accepting a moral constraint on the direct pursuit of one's greatest utility" (p. 161).

Hobbes presents a reply to the Foole. According to Hobbes, the Foole underestimates the risks involved in violating 'Covenant'. In his words:

> He [...] that breaketh his Covenant, and consequently declareth that he thinks he may with reason do so, cannot be received into any Society, that unite themselves for Peace and Defence, but by the errour of them that receive him; nor when he is received, be retained in it, without seeing the danger of their errour; which errours a man cannot reasonably reckon upon as the means of his security: and therefore if he be left, or cast out of Society, he perisheth; and if he live in Society, it is by the errours of other men, which he could not forsee, nor reckon up; and consequently against the reason of his preservation; (Hobbes, 1651/1991, p. 102)

Gauthier interprets Hobbes as saying that a person who is disposed to violate his covenants cannot be accepted as a party to beneficial arrangements by

those who are both rational and aware of his disposition, and therefore cannot rationally expect to reap the benefits that are available to people who are disposed to keep their covenants.[2] On this interpretation, "Hobbes moves the question from whether it be against reason, understood as utility-maximization, to keep one's agreement (given sufficient security of others keeping their agreements), to whether it be against reason to be *disposed* to keep one's agreement" (p. 162, my italics). Gauthier's own response to the Compliance Problem is an elaboration of this interpretation of Hobbes.

3 Constrained maximization and translucency

"The essential point in our argument", Gauthier writes, "is that one's disposition to choose affects the situations in which one may expect to find oneself" (p. 183). Gauthier argues it is not in a person's interest to be someone who is disposed to violate moral norms because others will refrain from cooperating with such a person. People do not want to cooperate with someone who cannot be trusted to respect the terms associated with cooperation; who will tell a lie, break a promise, etcetera, when it is in his interest to do so. The disposition to take self-benefiting opportunities is thus costly for a person. The disposition to comply with moral norms provided others do so as well, on the other hand, may by similar reasoning be expected to be beneficial. A person who internalizes moral norms as constraints on the direct pursuit of his interest may expect to be welcomed into cooperative arrangements, Gauthier claims. This disposition, which he calls *constrained maximization*, is therefore advantageous to have. More precisely, he claims constrained maximization is more advantageous than alternative dispositions, such that a rational utility-maximizer who does not yet have it would choose to have it. That implies, Gauthier argues, that constrained maximization is rational.

Gauthier's response to the Compliance Problem can be divided into two separate claims. The first is the claim that it is rational to internalize moral norms as constraints on the pursuit of self-interest. This is an empirical claim:

[2] There are of course other interpretations possible. To name one alternative, Hobbes has been interpreted as stating that agents cannot rationally expect violation of an agreement to be advantageous, even though it may actually be advantageous. While from a short-term point of view violating may be judged rational, potential long-term effects are so harmful—one may be thrown out of society!—that they outweigh any expected short-term benefit (cf. Gauthier, 1969; Skyrms, 1996). According to this interpretation, Hobbes denies that self-benefiting opportunities occur.

it is true if and only if it indeed serves a person's best interest to be a constrained maximizer. The second is the conceptual claim that if it is rational to be a constrained maximizer, the choices of a constrained maximizer are also rational. Gauthier requires this second claim to be able to say that a constrained maximizer can rationally choose morally when doing so is not in her best interest. This second claim has received most of the attention of critics.[3] I will, however, only investigate the first claim.

Given its central role in Gauthier's argument, I should start by saying somewhat more about the nature of a "disposition to choose". Gauthier is not very clear on it, and critics have interpreted the notion in various ways. Constrained maximization is sometimes interpreted as some sort of habit that, once properly adopted, causes the agent to comply in the relevant situations (e.g. Nelson, 1988). But as Den Hartogh (1999) points out, this does not fit Gauthier's insistence that the constrained maximizer chooses rationally when she complies with the norms that she has internalised. Gauthier writes that the constrained maximizer "is someone who takes her reasons for acting, not only directly from the utilities of possible outcomes she may bring about, but also from her plans and commitments" (1993, p. 186). We may initially describe constrained maximization as first and foremost a disposition to choose to cooperate, provided certain conditions are met.

By putting forward the argument for constrained maximization, Gauthier does not claim that it is always rational to choose morally. In the first place, Gauthier emphasises that it is not rational to cooperate when others are not inclined to do so, and as such let others take advantage of you. The constrained maximizer is conditionally cooperative: she avoids cooperating with persons who are not disposed to cooperate themselves. I take this to mean that constrained maximizers may sometimes violate moral norms.[4]

More fundamentally, Gauthier's argument for constrained maximization only yields the conclusion that it is rational to be moral if certain empirical conditions are met. First, one must find oneself in a society with a sufficient number of other constrained maximizers around to cooperate with. When amoral persons outnumber constrained maximizers, the expected costs of others taking advantage of one's constraints may outweigh their benefits.

---

[3] The idea is criticised by, among many others, David Copp (1991), Gregory Kavka (1995), John Broome (2001) and Derek Parfit (2001; 2011). It is more favourably discussed by Michael Thompson (2008) and Richard Holton (2009).

[4] In particular, they may sometimes, in the terms of Kavka perform defensive violations of moral norms, "violations undertaken to protect the agent from being taken advantage of by others who violate" (1995, p. 8).

Following Hume, Gauthier holds that if we "fall into a society of ruffians" we must not be constrained maximizers but "consult only the direct dictates of our own utilities" (p. 181).

Second, one must find oneself in a society in which the generally accepted norms approximate those that would be the object of agreement. The argument for constrained maximization is first of all meant to solve Gauthier's Compliance Problem: to show that when moral norms are the object of agreement, it is rational to internalise them as constraints on the pursuit of self-interest. The argument does not imply, however, that were we to find ourselves in a society in which an alternative set of norms is generally accepted, it is rational to internalise those norms that Gauthier identifies as moral. To the contrary, it may be used to defend the rationality of internalising these alternative norms: what matters most is not the exact content of the norms, but that being disposed to comply with them yields cooperative opportunities.[5] In what follows I will assume that this second condition is met and that the norms internalised by the constrained maximizer are moral norms.

The third condition is the one on which my investigation will focus, and which I will discuss in the remainder of this section. Why, it may be asked, must one choose to *be* a constrained maximizer rather than merely appear to be one? As Gauthier puts it himself:

> Is not the Foole's ultimate argument that the truly prudent person, the fully rational utility-maximizer, must seek to appear trustworthy, an upholder of his agreements? For then he will not be excluded from the co-operative arrangements of his fellows, but will be welcomed as a partner, while he awaits opportunities to benefit at their expense—and, preferably, without their knowledge, so that he may retain the guise of constraint and trustworthiness. (1986, p. 173)

For Gauthier's response to the Compliance Problem to succeed he must reject this possibility. That is to say, he must assume persons can recognise one another's true dispositions.

Gauthier hints at the possibility of solving this problem in his own theory by introducing an idealising assumption. Besides assuming that parties to an agreement are rational and have neither moral nor other-regarding preferences, he may assume their dispositions are *transparent*. Consequently,

---

[5] Gauthier (1986) relies on this when he uses the argument for constrained maximization to defend compliance with existing norms that approximate those that would be the object of agreement (see p. 168). He also realises the argument may, under different conditions, be used to defend compliance with norms that do not approximate his standard (see p. 179).

they can only gain the cooperation of others by adopting constrained maximization.

Such an assumption cannot be made, however, if the argument for rational compliance is to apply to actual agents. Remember, bargainers are assumed to be rational, amoral, and mutually disinterested to ensure that their agreement is not based on irrationalities, moral presuppositions, and contingent feelings. These assumptions are made to ensure *we* have reason to accept the hypothetical agreement given only our non-moral and self-regarding interests. Assuming transparency would have the opposite effect, however: showing that parties who are transparent choose constrained maximization *due to transparency* in no way shows that we, non-transparent creatures, have reason to do so. As Gauthier writes, "to assume transparency may seem to rob our argument of much of its interest… we shall have failed to show that under actual, or realistically possible, conditions, moral constraints are rational" (p. 174).

Gauthier therefore makes an alternative assumption which he does deem defensible. In his words, "[w]e may appeal instead to a more realistic *translucency*, supposing that persons are neither transparent nor opaque, so that their disposition to co-operate or not may be ascertained by others, not with certainty, but as more than mere guesswork" (p. 174). Three aspects of Gauthier's notion of translucency should be highlighted. First, for Gauthier, translucency is not an all-or-nothing matter, but comes in degrees. To assume persons are translucent is to assume that the probability of their dispositions being correctly identified by others is higher than the probability with which others would identify them had they resorted to mere guessing. I will refer to the probability that a person's disposition is correctly identified as that person's *degree of translucency*.

Second, translucency refers to persons as well as their observers. To say of a person that he is translucent to a certain degree is also to say that others have some ability to identify his disposition. To investigate how translucent people are we may thus look at how able they are at recognising such dispositions.

The third point concerns what Gauthier says about the object of translucency, the property that others can ascertain. Gauthier refers here to constrained maximization as a "disposition to co-operate". The idea of translucency implies that having the disposition amounts to more than merely being disposed to choose to cooperate or not in the appropriate circumstances. For persons to be able to detect whether others are constrained maximizers or

not before having interacted with them, there must be information about states associated with the disposition. Information about a person's intentions is an obvious candidate: a constrained maximizer sincerely intends to cooperate with respect to future self-benefiting opportunities.

It is worth emphasising that Gauthier is not the only contract theorist to make an assumption about people's ability to recognise the disposition of others. Indeed, Hobbes himself makes such an assumption when he says that persons who do not keep their covenant cannot rationally expect to be accepted by others as a member of society; these others must thus have some ability to identify whether persons are trustworthy or not.[6] More recently, Gregory Kavka (1995), another Hobbesian contract theorist, has argued that we should follow moral rules because others are likely to detect our violations and will subsequently avoid cooperating with us. Put differently, others will take such behaviour as indicative of an untrustworthy disposition.

I have argued that in order for Gauthier's argument for constrained maximization to yield that it is rational to be moral, three conditions must be met: there must be sufficiently many other constrained maximizers, the norms they accept must approximate those that would be the object of agreement, and their dispositions, as well as alternative dispositions, must be translucent. This reveals that whether it is rational to be moral is on Gauthier's view largely an empirical question. Or as he puts it, "[w]hat constrained maximization does is to provide for the *possibility* of morality" (1993, p. 188, my italics).

The above also clarifies that Gauthier does take the above three conditions to be either "actual, or realistically possible" (p. 174). Many critics have doubted whether this is the case for translucency. For example, Nelson writes that "Gauthier does nothing to convince us that some particular degree of translucency is "reasonably" attributed to us" (1988, p. 159), Buchanan calls it "ad hoc" and "a dubious sweeping empirical generalization" (1990, p. 240), and Franssens deems it "psychologically very implausible" (1994, p. 270).

The way in which Gauthier introduces the idea of translucency is indeed somewhat ad hoc. In fact, Gauthier introduces it in *Morals by Agreement* after Derek Parfit had convinced him that an earlier presentation of his argument implicitly depended on the empirically implausible assumption of transparency (see Gauthier, 1975; Parfit, 1984, pp. 18-19). It should also be acknowledged

[6] It is also worth noting in this regard that in the introduction to *Leviathan*, Hobbes describes how persons may "read" one another, and does so in terms remarkably similar to those of the social psychologists discussed in 4§3.2.

that Gauthier does very little to convince us that people are as translucent as his argument requires. He says his argument applies to "beings as translucent as we may reasonably consider ourselves to be" (p. 174), but he does not provide considerations for thinking that we are indeed to a certain degree translucent.

However, this does not mean that he is incorrect. It may be the case that people are translucent to such a degree that it is advantageous for agents to be constrained maximizers. I call the assumption that this is indeed so the *Translucency Assumption*. Being an empirical assumption, its plausibility should be considered in the light of empirical findings. In order to find out what we must look for, the remainder of this chapter discusses what must be the case for the Translucency Assumption to be met.

Before moving on, it is worth noting that for constrained maximization to be rational it is not sufficient that it is the most advantageous disposition to have. Rationality can only require agents to adopt constrained maximization if they can do so. Gauthier thinks this condition is satisfied. As he puts it:

> At the core of our rational capacity is the ability to engage in self-critical reflection. The fully rational being is able to reflect on his standard of deliberation, and to change that standard in the light of reflection. Thus we suppose it possible for persons, who may initially assume that it is rational [to have an amoral disposition] to reject it in favour of constrained maximization. (Gauthier, 1986, pp. 183-184)

Not everyone has been convinced by this view. Jean Hampton, for one, writes that "the idea that one could 'will' to be disposed to act as Gauthier describes is dubious" (1991, p.41). In response to this, it should be pointed out that Gauthier's argument does not require that internalising moral norms as constraints is a quick and easy process. It would not be a problem for the argument if doing so requires training or education.[7] While this brief response surely does not settle the issue, I take it to justify working on the assumption that persons can become constrained maximizers.

4 Constrained maximisation versus straightforward maximisation

Gauthier argues for the rationality of constrained maximisation by arguing that it would be rationally chosen over alternative dispositions. As he puts it, "the idea of a choice among dispositions to choose is a heuristic device to

[7] Situationists may challenge whether it is even possible to *have* a disposition such as constrained maximization. I leave this concern for another occasion.

express the underlying requirement, that a rational disposition to choose be utility-maximizing" (1986, p. 183). Whether constrained maximization is rational thus depends on the expected advantage of alternative dispositions. An important contender is what he calls *straightforward maximization*: the disposition to directly pursue one's interests, and thus to violate cooperative norms whenever it suits one. Gauthier argues for the rationality of constrained maximization by comparing it with this alternative. As this comparison provides insight into how translucent persons must be for the Translucency Assumption to be satisfied, I shall now describe this comparison. A formal presentation can be found in the appendix (§1).

Gauthier writes that in order to assess which of these two dispositions is more advantageous, we only need to consider situations in which they yield different choices. As we saw, the argument for constrained maximization is meant in the first place to show that it is rational to comply with an agreement to cooperate in mutually beneficial ways even when violating the agreement is individually more advantageous. We must thus concentrate on this type of situation, which is what I called a self-benefiting opportunity. Constrained maximizers and straightforward maximizers choose differently with regard to such situations. Straightforward maximizers defect, as this is per definition in one's interest. Constrained maximizers, on the other hand, comply with the norm to cooperate provided they expect the others involved in the activity or practice to do so as well. In the following, I will use 'cooperate' as a shorthand for complying with an agreement or norm to cooperate. [8]

To compare the expected utility of constrained and straightforward maximization, we need a formal representation of a self-benefiting opportunity. Gauthier chooses to rely on the familiar two-player Prisoner's Dilemma (henceforth PD) for this purpose (Table 1). Situations with this structure include a prospect for mutual cooperation: each is better off if both cooperate than if they do not cooperate. There is also a prospect for individually beneficial defection: in case the other cooperates, each is better off

[8] Following Gauthier, I assume that dispositions are categorical and do not come in degrees. That is not to deny the possibility of being someone who violates norms only a certain extent of the time rather than always when confronted with self-benefiting opportunities, or someone who cooperates only in a number of the interactions in which she believes her interaction partners to be trustworthy: these are simply alternative dispositions to straightforward and constrained maximization, of which we can construe infinitely many. I say more about this in §5.3.

by defecting. Constrained maximizers and straightforward maximizers choose differently in a situation with this structure:

> In such a situation, a straightforward maximizer chooses not to co-operate. A constrained maximizer chooses to co-operate if, given her estimate of whether or not her partner will choose to co-operate, her own expected utility is greater than the utility she would expected from the non-co-operative outcome (Gauthier, 1986, p. 170)

While the PD is just one example of a self-benefiting opportunity, I will follow Gauthier in using it as a test case to compare the expected utility of constrained and straightforward maximization.

Table 1: Prisoner's Dilemma, with $e > c > d > f$

		Future interaction partners: Cooperate	Future interaction partners: Defect
The agent: Cooperate		c, c mutual cooperation	f, e sucker's payoff, exploitation
The agent: Defect		e, f exploitation, sucker's payoff	d, d noncooperation

Self-benefiting opportunities in general and the PD in particular are examples of strategic interactions. Agents in such interactions must choose their strategy while taking into account what strategies their interaction partners may pursue. Constrained and straightforward maximizers choose differently in such interactions, and may thus be said to have alternative "disposition⌜s⌝ for strategic choice" (Gauthier, 1986, p. 183). It is important to distinguish such choices from the choice among dispositions that Gauthier uses as a 'test' of the rationality of constrained maximization.[9] Indeed, Gauthier does not even consider this to be a strategic choice:

> although the choice is about interaction, to make it is not to engage in interaction. Taking others' dispositions as fixed, the individual reasons parametrically to his own best disposition. (Gauthier, 1986, p. 171)

The choice among dispositions counts not as strategic, Gauthier says, because what dispositions others in the population have may be taken to be largely

[9] The phrase is Danielson's (1991).

independent of one's own disposition.[10] By taking the dispositions of others to be fixed, the choice among dispositions is what is called a *parametric* choice.

If one does not distinguish carefully between the global choice among dispositions and the local choices to which the disposition pertains, it may seem that straightforward maximization must be more advantageous than constrained maximization. But while defecting is the utility-maximizing choice in situations with the structure of the PD, it is not necessarily utility-maximizing to choose to be someone who defects in such situations. This is easiest to see if we assume persons are transparent. Whereas transparent straightforward maximizers will in each interaction end up with the punishment for defection, transparent constrained maximizers will get the reward for mutual cooperation whenever interacting with other constrained maximizers.

As I explained in the previous section, Gauthier does not assume that people are transparent, but that there is nevertheless a decent chance that their dispositions are identified correctly. In his comparison of constrained with straightforward maximization, two probabilities are of particular importance: the probability p that two constrained maximizers achieve mutual recognition, and the probability q that a straightforward maximizer is mistaken for a constrained maximizer. The more translucent constrained maximizers are, the higher the value of p; the more translucent straightforward maximizers are, the lower the value of q.

The Translucency Assumption states that people are to such a degree translucent that it is advantageous for agents to adopt constrained maximization. Whether this assumption is correct depends essentially on two things: what degree of translucency persons actually have, and on how high this degree must be if constrained maximization is to be the most advantageous disposition. Gauthier distinguishes two factors which affect the value of this latter threshold. First, it depends on the outcomes associated with self-benefiting opportunities. As we take the PD as a model for such situations, it depends on what values the agents assign to its four outcomes. Other things being equal, the larger the difference between the payoffs of cooperation and those of noncooperation, the lower the value of translucency required for

[10] It should be noted that Gauthier's reliance on a parametric choice model has been criticised extensively. Several theorists have reconstructed it as a strategic choice in a normal-form game (Bicchieri, 1993; Franssen, 1994) or as a competition between strategies in an evolutionary game theoretic model (Danielson, 1991).

constrained maximization to be more advantageous than straightforward maximization. On the other hand, when the difference between the payoffs of exploitation and those from cooperation increases, or the difference between the sucker's payoff and the payoffs of noncooperation, so does the translucency that persons must have for it to be rational to choose constrained maximization.

The second factor on which the value of the threshold depends has already been mentioned in the previous section: how many constrained maximizers are around to interact cooperatively with. In his comparison of constrained and straightforward maximization, Gauthier accounts for this factor by means of a probability variable r that an interaction partner is a constrained maximizer. He makes the assumption that everyone is either a constrained or a straightforward maximizer, in which case the probability that an interaction partner is a straightforward maximizer is $1 - r$. It is not hard to see that the degree of translucency required for constrained maximization decreases when the probability of interacting with a constrained maximizer increases.

Taking the probabilities p, q, and r and the payoff values together, the expected utility of the two dispositions can be compared. I do this in the appendix (§1). The resulting formula provides insight into how translucent persons must be for constrained maximization to be more advantageous than straightforward maximization. Moreover, we can derive certain necessary conditions from it. Particularly noteworthy is, as Maarten Franssen (1994) has shown, that whatever the payoffs and the probability r of facing a constrained maximizer, the probability $\sqrt{p}$ that a constrained maximizer is recognised must be higher than the probability q that a straightforward maximizer remains undetected (see appendix, §2). Moreover, Gauthier himself points out that with respect to a payoff distribution in which the differences between exploitation, cooperation, noncooperation and the sucker's payoff are equal, which he himself deems it defensible to suppose, the probability $\sqrt{p}$ must be more than *twice* the probability q (see appendix §3). These results show that people must be quite translucent for constrained maximization to have a higher expected utility than straightforward maximization.

5 Three challenges to the Translucency Assumption

Gauthier's claim that persons are to such a degree translucent that it is advantageous to be a constrained maximizer has been received rather critically. I structure my evaluation of the Translucency Assumption by concentrating on three important challenges that have been posed against it. This section introduces these challenges and explains how they will be addressed in the following chapters.

5.1 First challenge: Translucency is psychologically implausible

Many critics of Gauthier have expressed doubt about whether persons are as translucent as his argument for the rationality of constrained maximization requires. Some have gone further, expressing doubt about whether people are translucent *at all*. Ken Binmore (1993), for example, writes about the Translucency Assumption that "one cannot simply add 'idealizing assumptions' willy-nilly. People cannot see inside each other's heads and it is idle to examine models in which they can" (p. 138). As I mentioned in section 3, by assuming that persons are translucent, Gauthier assumes that the choices to which such dispositions pertain are preceded information that observers can pick up. I take Binmore to deny the existence of such information. This constitutes a first challenge to the Translucency Assumption: the whole idea that we can recognise each other's disposition is psychologically implausible.

This seems to be the sort of challenge that can be investigated on the basis of empirical findings. Studies may reveal whether persons can or cannot detect the dispositions of others. There is, however, a methodological difficulty. The dispositions of constrained and straightforward maximization are not examined in the empirical literature. As such, there are also no studies concerning the translucency of persons with these dispositions. This difficulty can be overcome by relating constrained maximization to the more familiar notion of trustworthiness. When a constrained maximizer decides to cooperate with another party, she provides her interaction partner with an opportunity to exploit her. However, she only makes herself vulnerable because she expects the other to cooperate as well; she *trusts* that the other party will not take advantage of her. In the context of a freely made decision to engage in a cooperative interaction with another, trust is only warranted when one has reason to expect this other person to be trustworthy. *Trustworthiness* is the property to be both competent and committed to do what one is trusted to do when one has the opportunity to betray this trust (McLeod, 2011). In the

context of commonly accepted norms of cooperation, it is thus the willingness to adhere to these norms and cooperate with others. It usually does not refer to an unconditional disposition to comply: it is the willingness to adhere to cooperative norms when one expects others to do so as well (Boone & Buck, 2003). Trustworthiness and constrained maximization are thus closely related.[11] I shall assume that constrained maximizers look for the property of trustworthiness in their interaction partners. As such, I will take translucency to refer to the probability that a person who is (un)trustworthy is recognised as such. Under this assumption, I will examine whether available empirical studies regarding our ability to detect trustworthiness reveal that persons are translucent or not.[12]

5.2 Second challenge: People are not sufficiently translucent for constrained maximization to be more advantageous than straightforward maximization

Most critics of the Translucency Assumption do not so much doubt that persons have some degree of translucency, but doubt that it is sufficient for Gauthier's argument to succeed.[13] In its strongest form, the challenge is that persons in general are insufficiently translucent for constrained maximization to be more advantageous than straightforward maximization. I will evaluate this challenge by examining what studies show about how likely it is that trustworthiness and untrustworthiness are detected by others.

Besides the strongest form of this challenge, it may also be posed more subtly by emphasising interpersonal differences in translucency. Translucency depends on properties that vary between individuals. For example, persons

[11] It is worth noting that in an early version of his argument Gauthier (1967) describes a type of agent that seems identical to the constrained maximizer and calls it the "the trustworthy man" (p. 473).

[12] The relation between untrustworthiness and straightforward maximization is less clear. While a straightforward maximizer is clearly untrustworthy—he agrees to cooperate but violates this agreement whenever it suits him—not every person whom we would describe as untrustworthy fits the profile of a straightforward maximizer. This does not need to pose a problem for the present investigation. Studies that concern untrustworthiness in general apply also to more specific types of untrustworthiness. That there are other types of untrustworthiness besides the straightforward maximizer's kind does, however, point at a limitation of Gauthier's analysis. There may be alternative untrustworthy dispositions to which constrained maximization must be compared. I return to this point when discussing the third challenge to the Translucency Assumption.

[13] As is shown in the appendix (§3), under an apparently plausible assumption about the benefits of cooperation and exploitation, constrained maximizers must be at least twice as likely to achieve mutual recognition than straightforward maximizers are to succeed in remaining hidden.

who are skilled at concealing their dispositions are less translucent than persons less skilled in this regard. On the basis of such considerations, it may be argued that there are persons who can keep their translucency so low that they are better off as straightforward than as constrained maximizers. Geoffrey -McCord has defended such a claim:

> Deceptive people will be careful to provide the requisite (though misleading) evidence [about their disposition] for those with whom they interact. They will develop winning smiles, travel with a glowing reputation, and cultivate an honest manner. Sadly, this sort of magic worked (without a ring of Gyges) all too frequently. Such people seem both translucent and trustworthy. When the deceptive have worked their magic, their companions will quite reasonably, given their information, misjudge character with regularity. (Sayre-McCord, 1991, p. 192)

According to Sayre-McCord, it is highly plausible that a substantial minority of agents is, due to their deceptive powers, better off as straightforward maximizer. For these 'deceptive people' it would thus not be rational to choose constrained maximization.

It is important to emphasise that the Translucency Assumption does not exclude there being interpersonal differences in translucency.[14] But as Gauthier wants to defend the rationality of morality, he is committed to the claim that constrained maximization is rational *for everyone*, provided certain background conditions are satisfied (see §3). The Translucency Assumption thus requires that even for those low in translucency it is advantageous to be a constrained maximizer.

It will be very difficult, if not impossible, to show that such deceptive persons may not exist. Psychological studies tend to focus on averages rather than on individual differences. But the tables can be turned. For us to have reason to think there are persons who are sufficiently deceptive to be better off as straightforward maximizers, there must be evidence of their existence. I will therefore consider whether there is reason to think that there are people for whom, due to their powers of deception, straightforward maximization is more advantageous than constrained maximization. If there is positive evidence, we have a serious challenge. If there is no evidence for such persons, we should not be too concerned about it. Like the first challenge, I shall consider this second challenge in the following chapter.

---

[14] Neither does the Translucency Assumption exclude there being *intra*personal differences in translucency: that persons vary in how translucent they are throughout their interactions. However, it does imply that the probability of being recognised averages in such a way that one should expect to be better off being a constrained maximizer than being a straightforward maximizer.

5.3 Third challenge: Reserved maximization is more advantageous than constrained maximization

The Translucency Assumption states that persons are to such a degree translucent that it is advantageous to adopt constrained maximization as one's disposition towards self-benefiting opportunities. Nevertheless, Gauthier compares constrained maximization with only one alternative disposition, straightforward maximization. This gap has invited many critics to imagine other dispositions that may do better than constrained maximization.

Of the infinite number of alternative dispositions that exist, there is one that may seem to do particularly well against constrained maximization once straightforward maximization be seen to fail due to translucency. This is what David Copp calls a reserved maximization:

> A reserved maximizer has exactly the disposition of a constrained maximizer, except that he will violate a requirement of a cooperative scheme whenever he has the opportunity to win the jackpot. He will take opportunities to make very great gains in utility, when the probability of detection is very low. For example, unlike a constrained maximizer, he may steal the money from a lost wallet, provided enough money is involved and provided he is quite sure he was not observed finding the wallet. A person might do better as a reserved maximizer than as a constrained maximizer. (Copp, 1991, p. 221)

As the reserved maximizer acts almost identically to the constrained maximizer, he may be expected to blend in and get similar cooperative opportunities. Add to this that he receives substantial benefits from occasional golden opportunities, and it seems that persons may well be better off as reserved than as constrained maximizers.

My evaluation of this third challenge shall concentrate on the question of how able we may expect persons to be at keeping up the appearance of being a constrained maximizer even though they are in fact reserved maximizers. Besides findings on our ability to distinguish trustworthy from untrustworthy others, studies on the risk of losing the trust of others will play a particularly important role in this discussion.

6 Conclusions

Contract theories must explain why people have reason to comply with principles that would be the object of agreement; or put differently, what reason they have for ensuring that their actions satisfy the contract test. I concentrated in this chapter on the Hobbesian contract theorist's answer to

this question, and in particular on that of Gauthier. Gauthier claims that persons have reason to adopt norms that would be the object of agreement as constraints on the direct pursuit of their interests because this in their interest: being a constrained maximizer enhances one's opportunities for beneficial cooperation. I have explained that this argument for constrained maximization requires that persons are to a certain degree translucent.

My investigation in the second part of this book concerns the plausibility of this argument. More precisely, it concerns the empirical plausibility of the assumption that people are translucent to such a degree that it is advantageous for agents to be constrained maximizers. I will evaluate this assumption by concentrating on three challenges that have been posed to it, which were introduced in the previous section. I will address the first two challenges in the following chapter, and the third challenge in Chapter 8. The final chapter examines the implications for Gauthier's argument for constrained maximization.

7

Translucency and the Irrationality of Straightforward Maximization

1 Introduction

Gauthier's assumption that people are translucent to such a degree that it is advantageous for agents to be constrained maximizers has been challenged, we saw in the previous chapter. Some critics doubt whether people are translucent at all, whereas others contest they are sufficiently translucent for adopting constraints to be more advantageous than not doing so. This chapter addresses these two challenges.

As I explained in the previous chapter (§5), I will use findings on trustworthiness to evaluate this Translucency Assumption. Constrained maximizers and straightforward maximizers differ qua trustworthiness. Constrained maximizers are trustworthy: they are willing to constrain the pursuit of their self-interest by moral norms when interacting with others, provided they expect them to do so as well. Straightforward maximizers, on the other hand, cannot be trusted to cooperate. They do not accept constraints on the pursuit of their interests, and thus violate agreements and norms whenever it suits them. Because of this difference in trustworthiness, if trustworthiness is so likely to be recognised that being trustworthy is more advantageous than being untrustworthy we also have reason to think that constrained maximization is more advantageous than straightforward maximization. Such a finding would thus support the Translucency Assumption. Conversely, if empirical findings show that being untrustworthy is more advantageous than being trustworthy, the Translucency Assumption should be rejected as being empirically implausible.

There is one issue that should be addressed before turning to empirical findings. Gauthier is sometimes interpreted as claiming that constrained

maximization is more advantageous than straightforward maximization with respect to *isolated* interactions; that is, interactions involving self-benefiting opportunities between individuals who are wholly unfamiliar with each other's pasts and who have no interaction future. I do not think this interpretation is correct. I take Gauthier to claim that constrained maximization is a more advantageous disposition overall, including interactions in which partners are better informed about one another. I shall nevertheless also consider how Gauthier's argument for the rationality of moral constraint fares if persons would expect only to have isolated interactions.

The investigation proceeds in three steps. The first step is to evaluate how translucent people are in isolated interactions. I shall rely here in particular on experiments regarding our ability to judge the trustworthiness of strangers. Such studies can be used to derive estimates of our translucency in isolated interaction. I shall argue that these studies confirm that the first challenge to the Translucency Assumption is mistaken: persons are translucent. However, they do not show that constrained maximization is more advantageous than straightforward maximization with respect to isolated interactions, and the second challenge thus stands.

The second step is to evaluate how translucent persons are when instead of being complete strangers to one another, interaction partners are better informed about one another either because they have observed each other on previous occasions or because they have access to the judgments of third parties. I call such interactions *informed interactions*. There are unfortunately no experimental studies available that directly provide estimates of how translucent people are under these conditions. I therefore take the findings on translucency in isolated interaction as a starting point and discuss, as much as possible on the basis of empirical findings, how a person's translucency changes when his observers become better informed. I shall argue that translucency may be expected to increase to such an extent that, provided one's interaction partners are usually constrained maximizers, constrained maximization has a higher expected advantage than straightforward maximization.

The third step of the investigation is to consider whether these two conditions are satisfied: may agents expect the majority of their interactions to be of the informed type and with constrained maximizers? I shall argue that given their having a certain degree of control over the nature of their interactions with others and given that they are to a certain degree translucent

to one another, these conditions may indeed be expected to be satisfied. This answers the second challenge.

2 Translucency in isolated interaction

This section investigates whether persons are to such a degree translucent in isolated interactions that it is better to be a constrained rather than a straightforward maximizer in such interactions. I start with explaining why there is good reason to think that persons in isolated interactions are translucent rather than opaque. I shall then review studies that provide information regarding *how* translucent persons are.

2.1 Signs of trustworthiness

For individuals to be translucent in isolated interaction it must be the case that their trustworthiness or untrustworthiness is systematically related to certain behaviours, and that others are able to recognise these behaviours as signs of such dispositions. From a theoretical point of view there is reason to think that this is the case. Trustworthy and untrustworthy individuals are disposed to choose differently in certain situations, which means not only that they behave differently but also that they differ psychologically. In particular, they have different intentions and are motivated by different concerns. While a trustworthy person is motivated by agreements and norms, the untrustworthy person is solely motivated by his personal interests. Many theorists hold that, given that we are a social species, we must have evolved abilities to detect such differences between trustworthy and untrustworthy individuals (Cosmides & Tooby, 1992; e.g. Frank, 1988; Sperber et al., 2010). The idea is that without such abilities cooperation would not be advantageous and thus would not have been established.

Are there behaviours that are systematically related to trustworthiness and untrustworthiness? It is commonly believed that nonverbal behaviour, and in particular emotional expressions, may contain signs of trustworthiness and untrustworthiness. Most notably, Robert Frank (1988) has argued that because emotions can sustain cooperative interaction under conditions in which it is advantageous to defect, emotional expressions have evolved as signs of trustworthiness. Put roughly, the idea is that, given that having emotional dispositions that motivate one to cooperate make one an attractive interaction partner, it is advantageous to reveal the possession of these emotional dispositions. This idea fits nicely with the fact that emotional

expressions are not fully under intentional control (Darwin, 1899/1998; Ekman, 2003; Porter & Ten Brinke, 2008). In part because we do not have full control over the muscles and other bodily processes associated with emotions, we do not always succeed in suppressing them. 'Emotional leakage', as Ekman calls it, occurs regularly. Similarly, we are not always able to intentionally activate muscles associated with an emotion when we do not genuinely have it. As Jon Elster points out, while some people may be able to cry when it suits them, "nobody can blush at will" (1998, p. 51).

Building on the view of Frank, Boone and Buck (2003) have argued that emotional expressiveness is itself a sign of trustworthiness. Because being emotionally expressive makes it harder to deceive, one can place more confidence in the honesty of an emotionally expressive person. Interestingly, Schug and colleagues (2010) have found that emotionally expressive individuals are indeed more willing to cooperate.

There is thus reason to think that trustworthiness and untrustworthiness are associated with certain behavioural signs. Findings in the booming field of social cognition suggest that people are quite able to detect these. Capacities for perceiving intentions and emotions in the behaviour of others develop in the first years of a child's life. Infants of only 12 months old are sensitive to the difference between an intention to help versus hinder and interpret the future behaviour of an actor on the basis of their earlier experiences (Kuhlmeier, Wynn, & Bloom, 2003), and at 15 months they can evaluate mental states of others (Onishi & Baillargeon, 2005). Before the child goes to school, she is able to recognise emotions in the facial expressions of others, even if these are people from different cultures (e.g. Ekman & Friesen, 1971), to accurately ascribe complex mental states to them (Frith & Frith, 2003), and more generally, as I described in Chapter 3 (§2.2), to understand perspectives different from her own. The child also develops what Sperber and colleagues call 'epistemic vigilance', the ability to distinguish information from misinformation, including the ability to distinguish trustworthy from untrustworthy individuals (Mascaro & Sperber, 2009; Sperber et al., 2010).

That we are born to be social is also confirmed by neuroscience. Neuroscientists have found such a large proportion of the brain to be involved in social interaction that they sometimes refer to it as 'the social brain' (e.g. Blakemore, 2008). An impressive range of findings suggests that people can typically identify the intentions and emotions of others accurately and quickly because observing others' actions and emotional expressions tends to

automatically activate the neural circuits of the observer associated with producing similar behaviour (e.g. Keysers & Gazzola, 2006).

Besides capacities for interpreting behavioural information, persons also have knowledge that may support the exercise of these capacities. The use of social cognitive capacities typically occurs in contexts of shared social practices and conventions, defining what it means to behave in a certain way (Gallagher & Hutto, 2007). Knowledge of these enables us to quickly detect when someone acts peculiarly, even when his behaviour deviates only very slightly from that of others. For example, a person only has to talk a little bit louder than people usually do for us to infer he is angry, deaf, or crazy. There are also conventions with regard to trustworthiness. When a trustworthy person has made a promise, for example, he will in due course inform the promisee if he turns out to be unable to fulfil it, and he will try to make up for it in some way or another. Importantly, conventions also prescribe how one should respond emotionally in specific situations. When a trustworthy person is unable to fulfil a promise, for instance, he is supposed to show some kind of regret. As emotional expressions are hard to fake and people are skilled at reading them, untrustworthy persons are particularly likely to have difficulty deceiving others when they are expected to express emotions they do not in fact have (Ross & Dumouchel, 2004). For trustworthy individuals this means that there will be situations in which they can 'test' whether their interaction partners are trustworthy.

I conclude that there is good reason to think that trustworthiness and untrustworthiness are associated with certain behaviours, and that persons have capacities to recognise such behaviours as signs of these dispositions. Put differently, there is reason to think persons are not opaque to one another, but translucent.

2.2 Translucency and cooperation

Now we know that persons have capacities to identify the (un)trustworthiness of others, the next question is *how* capable they are at this. Put differently, we need to know how translucent persons are in isolated interactions. I shall review three types of studies in this subsection and the following two: experiments regarding individuals' accuracy to predict whether others, with whom they have communicated verbally, will cooperate or not in strategic situations; experiments regarding the ability to do so on the basis of non-verbal information alone; and experiments concerning individuals' effectiveness to distinguish lies from truths.

While game theory predicts rational and self-interested individuals to defect in the single-shot Prisoner's Dilemma (henceforth PD) in public good dilemmas, years of experiments show that, with actual participants, there is in fact a significant baseline of cooperation. Analysing over 100 studies, Sally (1995) found that participants cooperated in almost 50% of the cases. If persons are translucent, we should expect that people are particularly inclined to cooperate with others when they are given the opportunity to observe them or interact with them. Numerous studies find exactly this (Bicchieri, 2002; Bohnet & Frey, 1999; Dawes, 1980; Ledyard, 1995; Sally, 1995). After a short-period of pre-play communication, cooperation rates increase well above the baseline rates. As Sally (1995) writes, "a 100 round prisoners' dilemma with discussion before each round would have 40% more cooperation than the same game with no discussion, and about 36% more cooperation than the same game with discussion every 10 trials" (p. 78). Once players are allowed to discuss the dilemma, the likelihood they will choose to cooperate increases dramatically.

Several explanations have been offered to explain this communication effect, but a recurring element in these explanations is that communication allows players to exchange information regarding their willingness to cooperate. Frank, Gilovich and Regan (1993) have provided more direct evidence of this. In their experiment, participants were asked to play a one-shot PD with two other participants. More important for our purposes, participants had to predict what strategies their partners would play. Before making their own choices and predictions about others, participants were provided with 30 minutes pre-play communication time, in which they could talk about whatever they wanted. Consistent with earlier findings, Frank and colleagues found the overall rate of cooperation to rise when subjects had more opportunity to communicate. More to the point, they found participants who could freely communicate to be able to predict the strategies of their interaction partner with accuracy: of the 198 judgments made, 76.2% were correct.

This accuracy is much higher than the accuracy that participants would have obtained by guessing, which Frank and colleagues take to be 64.8%.[1] Co-

[1] Frank and colleagues (1993) arrive at this number by calculating not how accurate an observer who predicted 50% cooperation and 50% defection would have been, which would yield a rate of 50%, but by calculating how accurate an observer would have been if he predicted in accordance with the actual prediction rates. Participants predicted cooperation 81.3% of the time while 73.7% in fact cooperated, and predicted defection 18.7% of the time whereas 26.3% in fact defected.

operators were more often correctly recognised as such than defectors: 130 out of 146 (89.0%) of the co-operators were predicted to cooperate, whereas 21 out of 52 (40.4%) of the defectors were predicted to defect. It should, however, not be concluded from this that people have no skill at detecting defectors: given that only 11% of the co-operators were predicted to defect, defectors were almost 4 times more likely to be predicted to defect than co-operators. Similarly, cooperators were much more likely to be predicted to cooperate than defectors: 89.0% versus 59.6%. The study thus finds that cooperators were much more likely to be predicted to cooperate than defectors, and defectors were much more likely to be predicted to defect than cooperators. A person's disposition to choose thus appears to have increased the likelihood of being detected as having that disposition, which is exactly what it means for persons to be translucent.[2]

Interestingly, when participants were explicitly prohibited from making promises or agreements in the period of pre-play communication, their predictions were substantially less accurate. Does this not conflict with the idea that persons are translucent? There is reason to think it does not. As Cristina Bicchieri (2002; 2006) points out, when persons trust another to cooperate they expect the other to comply with a norm of cooperation. Communication may thus increase cooperation *provided* a norm of cooperation has been activated. As Bicchieri puts it, "the effect of discussion on cooperation rates might precisely be due to the fact that discussing the dilemma often involves an exchange of pledges and promises, and the very act of promising focuses subjects on a norm of promise-keeping, as well as that it fosters expectations that a sufficient number of subjects will fulfill their promises" (2006, p. 148). But if persons are explicitly forbidden to make promises or agreements, as was the case in this alternative condition of the study by Frank and colleagues, a norm of cooperation may not be activated. In that case, whether a person cooperates or not does not depend on his or her trustworthiness, and the finding would thus not concern translucency.

Critics of Frank et al (1993) have pointed out that the obtained accuracy in defection prediction may have occurred simply because some participants announced they would defect (e.g. Ockenfels & Selten, 2000). In her replication of the study Brosig (2002) therefore excluded interactions in which such statements were made. Without these, results were in line with those of Frank

[2] Participants also had a sense of when they were most likely to be accurate: of the 160 predictions made with a confidence level of 75 or higher, 80% were correct, as compared with only 60.5% of the 38 predictions made with a confidence level below 75.

et al: 66.7% of predictions were correct, with 85.7% of the cooperators being detected and 31.1% of the defectors.[3] The slightly lower accuracy may be explained by the fact that in Brosig's study participants had only 10 minutes of pre-play communication instead of 30 minutes.

2.3 Judging trustworthiness on the basis of nonverbal information

The above studies suggest that if persons communicate verbally before deciding whether to trust one another in a strategic situation, they are not unlikely to correctly recognise their interaction partner's disposition. But what if communication is only non-verbal? Interactions in which we do not have an opportunity to communicate verbally are common, and may also include self-benefiting opportunities. The idea that one's emotional expressions tend to reveal one's dispositions suggests that (un)trustworthiness may be communicated non-verbally in such interactions. Several findings support this.

Persons tend to form judgments of trustworthiness quickly and can do so on the basis of non-verbal information alone. Willis and Todorov (2006) found that persons can form judgments of trustworthiness of pictured individuals after being exposed to their behaviour for only 100 milliseconds. Interestingly, such judgments based on 'thin-slices' of behaviour correlate highly with judgments made in the absence of time constraints. Willis and Todorov take this to show that judgments of trustworthiness are formed quickly and "are created effortlessly on-line from minimal information" (p. 597). It appears that the amygdala, a neural structure involved among other things in emotional learning, responds automatically to facial properties associated with untrustworthiness (Engell, Haxby, & Todorov, 2007). The finding that judgments of trustworthiness can be arrived at very quickly may explain why cooperation rates in public good experiments have been found to increase after minimal non-verbal contact, such as activities like mutual gazing and gentle touching (Kurzban, 2001), or short periods of eye contact (Bohnet & Frey, 1999).

Are such quick-and-dirty judgments of trustworthiness accurate? There is good reason to think that we do indeed have some sensitivity for nonverbal signs of untrustworthiness. Vanneste, Verplaetse, Van Hiel and Braeckman (2007) have found that the faces of untrustworthy persons attract more attention than the faces of trustworthy persons. Vanneste and colleagues

[3] Jeannette Brosig informed me about the specific percentages for cooperation and defection by email.

measured the automatic attention of persons by means of a so-called dot probe classification task. In such a task, participants have to classify a probe as quickly and accurately as possible after being shown a stimulus image. The idea is that response time is a measure of the attention allocated to the presented image: if participants respond more slowly after one image than after another, this is taken to show that the former was given more attention. In the task set by Vanneste and colleagues, the stimuli used were facial pictures of persons who had been photographed by a webcam while playing a PD. Three types of pictures had been made of these persons: while displaying a neutral expression before playing the game, while choosing either to cooperate or defect in a practice round without any payoffs, and while making this choice in the actual game. Vanneste and colleagues found that participants responded significantly more slowly to the probe after seeing a facial picture of a person defecting in the actual game. They take this to indicate that untrustworthy faces attract our attention more than trustworthy faces, and that we thus have a sensitivity for nonverbal signs of untrustworthiness.

That such information can also inform our judgments is shown convincingly by another study by these researchers. Verplaetse, Vanneste and Braekman (2007) asked participants to predict whether pictured persons would cooperate or defect, again making use of the three types of pictures just mentioned. With regard to pictures of persons who made actual choices, Verplaetse and colleagues found participants to be surprisingly accurate: participants correctly classified 59% of the co-operators and 66% of the defectors, achieving an overall accuracy of 63%. Interestingly, participants' accuracy in classifying individuals who made choices in a practice round was no better than chance. As the researchers point out, given that we may expect persons to be less emotionally aroused in a practice round than when they play the game for real stakes, this supports the idea that emotional expressions provide information about trustworthiness.[4]

The above studies indicate that before persons make a choice to cooperate or not, they may unintentionally reveal their intention to others through nonverbal behaviours. But what if persons have not yet formed an intention with regard to a specific choice situation? There is evidence that even outside

[4] There may also be more stable facial cues of (un)trustworthiness. In an intriguing study, Stirrat and Perrett (2010) found not only that men with wide faces are significantly more likely to exploit their interaction partners in trust games than men with narrow faces, but also that men with wide faces are judged as less trustworthy than men with narrow faces. Similarly, there is some evidence that personality types such as psychopathy are associated with morphological cues (Holtzman, 2011).

of such situations, persons may unintentionally reveal information about how they would choose in them. Several studies that make use of the earlier mentioned thin-slices methodology report that we have some ability to detect altruism, a disposition that we may expect to be related with trustworthiness.[5] In one such study (Oda, Yamagata, Yabiku, & Matsumoto-Oda, 2009), participants were presented with 30-second silent videos of persons who the researchers had identified as either altruists or non-altruists on the basis of a questionnaire. It was found that participants tended to judge altruists as more altruistic than non-altruists. In another study (Fetchenhauer, Groothuis, & Pradel, 2010), participants had to predict the contributions of another person in a Dictator Game on the basis of a 20-second silent movie. The researchers found that participants performed significantly better than expected by chance, and that predicted contributions and actual contributions correlated. Importantly, as the behaviour that was videotaped took place in a context different from the one in which the observed persons made their choices, these studies suggest that there are stable nonverbal signs of altruism, in the sense that they may occur independently of altruistic choices.

Similarly, several thin-slices studies provide evidence that persons are sensitive to signs of personality types that bear some resemblance to the disposition of straightforward maximization (Back, Schmukle, & Egloff, 2010; Fowler, Lilienfeld, & Patrick, 2009; Holtzman, 2011). I am referring to the personality types narcissism, machiavellianism, and psychopathy, personalities that are sometimes referred to as 'the dark triad'. These three personality types are characterised by a disposition to deceive and manipulate other persons for their own gain and a disregard for morality. In one study on the ability to detect such personalities, participants were presented with 5 to 20 second segments of recordings of maximum-security inmates (Fowler et al., 2009). The segments involved several minutes of uninterrupted speech by the inmates that did not concern illegal or delinquent acts and included either both video and sound, only video, or only sound. Participants were asked to rate the recorded persons on several properties, amongst which were the crucial diagnostic properties of psychopaths. As the inmates had on an earlier occasion been scored on several measures for psychopathy, researchers could investigate how able people are at detecting the psychopaths among these inmates. Interestingly, they found that the thin-slice psychopathy ratings made by participants correlated with the earlier test scores. We apparently

[5] Whereas trustworthiness can be thought of as a conditional willingness to cooperate, altruism is sometimes thought of as an unconditional willingness to cooperate.

have some ability to detect psychopathic features. Given that the video segments did not involve relevant verbal talk and that correlations were highest for video-only fragments, the study provides further support for the existence of nonverbal signs of dispositions.

Of course, our interest is not in whether persons can distinguish altruists from non-altruists or psychopaths from non-psychopaths. The above thin-slices studies reveal, however, that we are able to detect complex psychological characteristics on the basis of nonverbal information alone. Add to this that there is some resemblance between the above dispositions and the dispositions of trustworthiness and untrustworthiness, and it is plausible that these dispositions are also associated with stable signs that observers may pick up on.[6] They thus provide further support that a person's (un)trustworthiness can be recognised on the basis of nonverbal behaviour, and in particular suggest that such a disposition may be recognised outside of self-benefiting opportunities, when the person has not yet formed an intention to cooperate or to defect.

2.4 Lie detection

I have presented above different types of studies that support the idea that persons are translucent. There is, however, also a line of research that has been taken to be inconsistent with this idea, namely studies on lie detection (cf. Ockenfels & Selten, 2000).

In a standard lie detection experiment, participants are confronted with a series of individuals recorded on video making statements that are either true or false. The individuals are usually making these statements to a third person, who may be either responsive or passive. Ordinarily, half of the messages a participant encounters are truths and the other half lies such that a guess has a 50% chance of being correct. Reviews from the literature conclude that we are not very accurate lie detectors. A recent meta-study on the basis of 206 documents with over 24,000 different judges finds that individual studies almost always fall within the range of 45% to 65%, with a mean accuracy of 54% correct lie-truth judgments (Bond & DePaulo, 2006). In other words, we do only slightly better than may be expected by chance.

---

[6] It should be noted that the finding that cooperation rates increase also when communication is nonverbal does not need to conflict with the idea that norm activation plays a role in explaining why participants cooperated, as norms can also be activated nonverbally. As Bicchieri (2006, p. 148) writes: "More often than not the activation process is unconscious; it does not involve much thinking or even a choice on the part of subjects" (p. 148).

In a follow-up study, the researchers examined whether there were individual differences in the ability to detect deception. Somewhat surprisingly, they found no evidence for individual differences in the ability to detect lies: people appear to be about equally able judges. What they did find were substantial differences in the ability to deceive. Not only are some people significantly less likely and others significantly more likely to be detected than people on average (the standard deviation is 5.5%), some are also more likely to be judged as honest and others as dishonest *regardless* of the veracity of their statements (the standard deviation is 11.6%). Put differently, for some people the probability of their lies being detected is well below the 54%, while for others it is well above it.

While these findings are not what one would expect if people were highly translucent, they do not contradict the idea that persons are moderately translucent in isolated interactions. For one, while lying is associated with untrustworthiness, untrustworthiness is a broader disposition; even if an untrustworthy person's lies are undetected as such, he may be detected as untrustworthy on the basis of other signs. What is more, there are several reasons to think that lie detection in actual interactions would be more accurate than in lie detection studies. To start with, the meta-study just mentioned finds that only a little more familiarity with a liar increases accuracy significantly (Bond & DePaulo, 2006). Also, when lying occurs in conversation and the liar may have to answer questions, the cognitive load of the liar may be expected to increase substantially (Vrij, Edward, & Bull, 2001; Vrij et al., 2008). Indeed, lie researcher Aldert Vrij and his colleagues have argued that increasing the cognitive load of another is, in isolated instances, the single most effective way of finding out whether another is speaking the truth (Vrij, Granhag, & Porter, 2011a).

Finally, and most importantly, lies in real interactions should typically involve larger stakes, and this should make them easier to detect. As Ekman and colleagues (1999) point out, as there are little to no stakes for the videotaped liars in lie detection experiments, emotions that tend to betray lies—fear, guilt, or excitement, for example—may not have been aroused. There is evidence that when the stakes of lies increase, so does our accuracy in detecting them (Bond & DePaulo, 2006; Frank & Ekman, 1997; Porter & Ten Brinke, 2011). This means that even though our ability to detect everyday lies is limited, there is reason to think we would do better with regard to the high-stake lies associated with self-benefiting opportunities. That is not to say, however, that we are *good* at lie detecting: even when the circumstances are

more favourable than those of participants in lie detection experiments, we are average lie detectors at best.

2.5 Evaluating the Translucency Assumption for isolated interaction

The studies reviewed support the view that people are translucent in isolated interactions. Indeed, if we take the isolated interactions of participants who must predict a stranger's strategy in a Prisoner's Dilemma as representative of isolated interactions in general, we may conclude persons are *moderately* translucent. The probability that a person's willingness to cooperate is recognised is not high, but substantially above what we should expect had they resorted to guesswork. As far as I know, there are no studies that contradict this.[7] I shall now consider what these findings mean for Gauthier's Translucency Assumption.

I should start by mentioning one additional assumption that must be made in order to use the findings for this purpose. Several of the above studies suggest that part of the explanation why trustworthiness and untrustworthiness can be distinguished is that they are associated with different nonverbal signs, and in particular with different emotional expressions. To take these findings to be indicative of the translucency of constrained and straightforward maximization, it must be assumed these dispositions are similarly related to certain emotional expressions. I thus assume that in so far as trustworthiness includes an emotional component due to which it may be distinguished from untrustworthiness, constrained maximization does as well. It may be worth noting that this is not incompatible with how Gauthier conceives the disposition, as he writes that "each has reason to prefer that everyone be affectively engaged by compliance, so that the familiar feelings of respect and resentment, of self-respect and

[7] There is only one study I know of that has been proposed as evidence against the idea that persons are translucent. Ockenfels & Selten (2000) have found that participants who do a bargaining game with other participants are not good at detecting the initial bargaining position of their bargaining partner. They take this to count against the view that persons are translucent in isolated interactions: apparently, people are able to keep this information from others. I disagree with this interpretation. Translucency has to do with our ability to identify the trustworthiness or the willingness to cooperate of others. That persons are able to hide certain information from others that has, as far as I can see, nothing to do with cooperation, does not count against the idea that people are translucent.

guilt, are linked appropriately with the fair and unfair behavior of others and oneself" (1986, p. 266).[8]

The above empirical studies show that persons are as a matter of fact translucent, and thus defeat the first challenge that was posed against the Translucency Assumption; there is nothing psychologically implausible about persons being translucent. But what about the second challenge? Do the findings suggest people are translucent to such a degree that, with respect to isolated interactions, adopting moral constraint is more advantageous than being a straightforward maximizer?

To answer this question, I draw on Gauthier's analysis that was introduced in the previous chapter (§6.4). Gauthier holds that to compare the expected advantage of constrained and straightforward maximization we need only consider self-benefiting opportunities. The simplest type of self-benefiting opportunity, and the one Gauthier uses, is the Prisoner's Dilemma. Gauthier defines constrained maximization as the disposition to cooperate in such dilemmas if one believes one's interaction partner is similarly disposed. Straightforward maximizers, on the other hand, defect. Like Gauthier, I also make the assumption that interaction partners that one faces are either constrained or straightforward maximizers themselves.[9]

As I explained in the previous chapter, which disposition has the higher expected utility depends on the degree to which oneself and others are translucent. Following Gauthier, we describe persons' degree of translucency by means of two probabilities: the probability $\sqrt{p}$ that a constrained maximizer is recognised by other constrained maximizers, and the probability q that

[8] Den Hartogh (1993) argues that if translucency were to have an emotional basis, this may pose a problem for Gauthier. As I have mentioned before, Gauthier's argument for constrained maximization is meant to show that a fully rational person without moral or other-regarding preferences would adopt moral constraint. Den Hartogh points out that as such persons do not have similar moral sentiments as we have, they may also not be similarly translucent. This in turn means, he argues, that they may not be sufficiently translucent for constrained maximization to be rational for them. I have two short responses to this objection. First, if it is correct, it poses no problem to the present investigation: I concentrate on the question of whether for actual persons, who usually do have such sentiments, it is advantageous to be constrained maximizers. Second, it is not evidently correct: the standpoint of a fully rational person is first and foremost a heuristic, and if such a person can choose a disposition such as constrained maximization I do not see why he may not also choose (or develop) associated sentiments that are required for the disposition to be advantageous. This seems to be Gauthier's own view (see 1986, p. 266).

[9] This idealising assumption is not fully innocent as the introduction of additional dispositions in the population may change matters substantially. In particular, if the population would include a substantial proportion of *unconditional* cooperators this could benefit straightforward maximization substantially.

constrained maximizers mistake a straightforward maximizer for a constrained maximizer.[10]

How much translucency is required for constrained maximization to be more advantageous than straightforward maximization depends on two factors. First, it depends on the payoffs associated with the different outcomes of the PD (Table 1).[11] Second, it depends on the probability r of one's interaction partner being a constrained maximizer. If the probability of interacting with a trustworthy person increases, so does the probability of achieving mutual cooperation, while the probability of being exploited decreases.

Table 1: Prisoner's Dilemma, with $e > c > d > f$

		Future interaction partners: Cooperate	Future interaction partners: Defect
The agent: Cooperate		c, c mutual cooperation	f, e sucker's payoff, exploitation
The agent: Defect		e, f exploitation, sucker's payoff	d, d noncooperation

The expected utility of constrained maximization and straightforward maximization can be calculated by combining the probabilities $\sqrt{p}$, q and r with the payoff paramaters. As I show in the appendix (§1), constrained maximization has a higher expected advantage than straightforward maximization if and only if:

$$(e - d)rq < (c + d - e - f)rp + (e + f - 2d)r\sqrt{p} + (d - f)rq + (f - d)q$$

As I noted in the previous chapter (§4), it can be shown from this that several conditions must be met for constrained maximization to be more advantageous

[10] Gauthier uses the probability p that two constrained maximizers attain mutual recognition. In order to also allow for the possibility of unilateral recognition among constrained maximizers, I follow Franssen (1994) in using $\sqrt{p}$ for the probability that one constrained maximizer is recognised by another constrained maximizer as such.

[11] Here and in the rest of this investigation I assume that payoffs represent only self-regarding interests (which is how self-interest is commonly defined (Shaver, 2010)). Whether a person has a self-benefiting opportunity is thus not affected by the prosocial or moral sentiments she may have.

than straightforward maximization. A first condition is that $\sqrt{p}$ must be higher than q (see the appendix, §2, for a proof). Do the findings presented in this section provide reason to think this is the case?

The studies on prediction accuracy in PDs can be used to estimate the values of $\sqrt{p}$ and q. One option is to base these estimates on the overall detection rates that these studies report. As we saw, they find detection rates of 76% (Frank et al., 1993), 67% (Brosig, 2002) and 63% (Verplaetse et al., 2007). Taking the average of these values and assuming that the probability of being detected does not vary with the agent's disposition or his observer's, we get 0.69 as the probability $\sqrt{p}$ that a constrained maximizer is recognised as such by other constrained maximizers, and 0.31 as the probability q that a straightforward maximizer is *not* detected by constrained maximizers. With these values, the first condition is evidently satisfied.

Overall detection rate may, however, not provide the best basis for our estimates. Gauthier presumably distinguishes the probability that constrained maximizers recognise a constrained maximizer ($\sqrt{p}$) and the probability that they recognise a straightforward maximizer ($1 - q$) to allow for the possibility that constrained maximizers and straightforward maximizers are not equally likely to be detected. In support of this distinction, the above studies find different detection rates for cooperators and defectors. Cooperators were detected 89% (Frank et al., 1993), 86% (Brosig, 2002), and 59% (Verplaetse et al., 2007) of the time, which gives an (unweighted) average detection rate of 78%. With regard to the detection rate of defectors, these studies respectively report values of 40%, 31% and 66%, giving an average detection rate of 46%. Assuming still that the probability of being detected does not vary with the disposition of observers, we can take the average rate by which cooperators were detected as an estimate of the probability $\sqrt{p}$ that a constrained maximizer is detected by other constrained maximizers, and the average rate by which defectors were *not* detected as an estimate of the probability q that a straightforward maximizer is not detected by constrained maximizers, meaning that we get an estimate for $\sqrt{p}$ of 0.78 and an estimate for q of 0.54. Again, with these values the first condition is satisfied.

While these are comforting results, this first condition being satisfied is with respect to most payoff distributions insufficient for constrained maximization to trump straightforward maximization. It is also insufficient with respect to the type of distribution that Gauthier concentrates on, displayed in Table 2. I call this a *symmetric* PD because the gain of cooperation

over noncooperation, the gain of exploitation over cooperation, and the loss of the sucker's payoff with respect to noncooperation are equal. In order for constrained maximization to be more advantageous than straightforward maximization with respect to such interactions, $\sqrt{p}$ must be at least *twice* as high than q (see the appendix, §3, for a proof).

Table 2: Prisoner's Dilemma, with $e > c > d > f$ and $e - c = c - d = d - f$

		Future interaction partners: Cooperate	Future interaction partners: Defect
The agent	Cooperate	2, 2	0, 3
	Defect	3, 0	1, 1

While this condition would be satisfied when we base estimates of $\sqrt{p}$ and q on the average overall accuracy rates reported by the above studies, this clearly is not the case when we split between detecting cooperators and detecting defectors. The most optimistic findings, which are those of Frank and colleagues (Frank et al., 1993), provide an estimate of 0.89 for $\sqrt{p}$ and 0.60 for q. These values are insufficient to make constrained maximization more advantageous than straightforward maximization with respect to symmetric PD's.

There are of course many other payoff distributions for which this degree of translucency is sufficient for constrained maximization to be more advantageous than straightforward maximization. This is for instance the case for distributions in which the gain from cooperation over noncooperation is substantially larger than the gain from defection over co-operation or the loss from noncooperation to exploitation.[12] However, Gauthier argues that, "on the whole, there is no reason that the typical gain from defection over co-operation would be either greater or smaller than the typical gain from co-operation over non-co-operation, and in turn no reason that the latter gain would be greater or smaller than the typical loss from non-co-operation to exploitation" (p. 177). Were we to follow Gauthier on this, it must be concluded that empirical findings suggest that for isolated interaction

[12] For example, under the condition that one is sure to face a constrained maximizer and r is thus 1, constrained maximization is more advantageous than straightforward maximization once the payoff of mutual cooperation is increased from 2 to 2.5.

constrained maximization is *not* more advantageous than straightforward maximization.

3 Translucency in informed interaction

When comparing the expected utility of constrained and straightforward maximization we should not only consider isolated interactions. Self-benefiting opportunities occur just as well in contexts in which persons are better informed about each other; between persons who have interacted with each other before, or who are informed about one another through third parties. In the present section I will discuss whether the Translucency Assumption is plausible with respect to such *informed* interactions: whether persons are sufficiently translucent in such interactions for constrained maximization to be more advantageous than straightforward maximization with respect to them. Whether constrained maximization is more advantageous *overall*, taking into account both informed and isolated interactions, will be considered in the next section.

For methodological reasons my approach must be different from that in the previous section. Contrary to isolated interaction, there are to my knowledge no studies that provide suitable estimates of persons' translucency in informed interaction. My approach will therefore be to take the results from the previous section as a starting point, and discuss, as much as possible on the basis of empirical findings, how these should be expected to change in informed interaction. I shall first discuss how a person's translucency is affected if his observers possess information from previous observations, and then how it is affected when they have access to judgments from third parties.

3.1 Translucency and previous observations

What happens to a person's translucency when his interaction partners have observed him on previous occasions? By having had more opportunity to observe his behaviour, such partners are more likely to have a correct judgment about his (un)trustworthiness than isolated interaction partners. With increased opportunity to observe, observers should become better at detecting signs. An example of this phenomenon is found in the literature on lie detection. When participants are provided with a short baseline exposure to the person whose statements they have to judge, their average accuracy increases by 4 percentage points (Bond & DePaulo, 2006). Moreover, simply because the number of signs of (un)trustworthiness that a person displays

increases with time, observers with more opportunity to observe tend to witness a higher number of signs.

Among the signs that such partners may detect are choices that a person has made with regard to past self-benefiting opportunities. Whenever a person chooses to cooperate or defect with respect to such a situation it may be detected by others. Sometimes such choices are detected instantaneously. This occurs, for example, when a person is observed either to act as she promised or not when doing so is not in her interest.

When a choice with respect to a self-benefiting opportunity is not observed directly, there is yet a chance of it being detected later in time. Lying is again a case in point. As several lie detection researchers have emphasised, detecting lies over time works differently from detecting them on the spot (Park, Levine, McCornack, & Morrison, 2002). Detective fiction illustrates this. It is a common theme in such works that, while the plot thickens, a prima facie trustworthy character turns out to have been lying. Even though his lies were undetected when produced, they were detected eventually due to inconsistencies with other information. As Park and colleagues (2002) point out, detective fiction teaches us several lessons about the nature of lie detection. The first is about the kind of information relevant for detecting a lie. Apart from directly observed behaviour, "people often rely on information from third parties, the consistency of statements with prior knowledge, the consistency of messages with physical evidence, or confessions when rendering judgments about the veracity of others' messages" (p. 145). The second is about time. Detecting a lie often occurs hours, days, weeks or even months after the lie was uttered.

This point applies to self-benefiting opportunities in general. A person's interaction partners may learn about her past choices with respect to such situations. The likelihood of this occurring is most likely for partners who were involved in those earlier interactions, but they may also learn about it from third parties.

Being reliable signs, a person's detected choices may affect his translucency substantially. This is certainly the case in a world in which everyone is either a constrained or a straightforward maximizer: if someone complies when it is in his interest to violate, he is a constrained maximizer; if he violates, he is a straightforward maximizer. Of course, it is unlikely that in reality everyone can be described as having either one of these dispositions—there are likely to be many people who often comply with norms and sometimes do not. Because of that, a person's choice to cooperate or to defect

in a self-benefiting opportunity is not a 100% reliable sign of what that person will do with regard to future self-benefiting opportunities. However, it has been found that people *do* in general take such choices to be highly reliable signs of trustworthiness. When a person has been caught lying or cheating, people tend to take this to reveal he is disposed to do so—to reveal that he is untrustworthy (Bond & DePaulo, 2006; O'Sullivan, 2003).[13] In line with this, studies with repeated Prisoner's Dilemmas find that persons are often willing to cooperate with other with whom they have no interaction history, but instantly lose their trust in their partner when he cheats (Gibson, Bottom, & Murnighan, 1999; Schweitzer, Hershey, & Bradlow, 2006).

There is reason to think this effect is stronger for violating than for cooperating. Conventional wisdom says that violations of trust have a much larger influence on the extent to which a person is trusted than behaviours consistent with trustworthiness. As the psychologist Paul Slovic writes:

> One of the most fundamental qualities of trust has been known for ages. Trust is fragile. It is typically created rather slowly, but it can be destroyed in an instant—by a single mishap or mistake. (Slovic, 1999, p. 697)

This asymmetry principle, as Slovic calls it, has been corroborated by experimental studies (e.g. Cvetkovich, Siegrist, Murray, & Tragesser, 2002; Slovic, 1993). These studies find that information suggestive of untrustworthiness influences trustworthiness judgments more than information suggestive of trustworthiness. It has been argued there is some rationale for this asymmetry. Violations are more informative of a person's trustworthiness than compliance (Poortinga & Pidgeon, 2004). For example, opportunists who only violate when there is much to gain will also cooperate in most cases, but are not trustworthy.

I conclude that, in comparison to isolated interactions, persons may be expected to be more translucent when their interaction partners have observed them on previous occasions. Their interaction partners are more likely to have detected signs of (un)trustworthiness, including previous choices with regard to self-benefiting opportunities. The trust-asymmetry principle states that violations tend to have a particularly large influence on the judgments of interaction partners. This suggests that if a straightforward maximizer's

[13] O'Sullivan (2003) describes this inference as an instance of the fundamental attribution error: the fact that a person does something once in a given situation does not mean he is disposed to do it. Nevertheless, as people tend to reason in this way it is something that must be taken into account when calculating the expected utility of a disposition.

interaction partners know about a transgression of his, they are likely to judge him to be untrustworthy.

3.2 Translucency and third party judgments

When deciding whether to trust a given person or not, we often have access to the judgment of third parties. This includes the judgments of our friends, but also those of strangers, as when we consider the facial expressions of others when confronted with a person who acts unexpectedly in public. What happens to a person's translucency if his observers have access to third party judgments?

Perhaps unsurprisingly, third party judgments of a person's (un)trustworthiness tend to influence how we judge that person ourselves. Burt and Knez (1996) report that gossip about a person's (un)trustworthiness substantially affects individuals' judgments of trustworthiness. Studying how individuals respond to information about their interaction partner's reputation in a public good game, Milinski and colleagues (2002) found that participants who learn their interaction partner has defected in a previous interaction cooperate only 20% of the time whereas 60% cooperates if they learn he has cooperated in previous interactions. Again, there is reason to think information indicative of untrustworthiness is more influential than information indicative of trustworthiness. Just as untrustworthy facial expressions tend to attract our attention (Vanneste et al., 2007), we are more attentive to information indicative of untrustworthiness and more likely to distribute it (Burt & Knez, 1996). Furthermore, it has been reported that receivers of third party information trust negative information more than positive information (White, Pahl, Buehner, & Haye, 2003).

A person's reputation is a special case of third party judgments to which we often have access. It is commonly believed that the trust-asymmetry principle applies here as well; for example, the investor Warren Buffet supposedly said, "it takes 20 years to build a reputation and five minutes to ruin it". A dramatic finding in this regard concerns the reputational effect of criminal accusation or convictions. Studying 369 drug convictions, Lott (1992) found that being convicted substantially decreases (legitimate) earnings when returning to the labour force. For example, a person with a pre-sentence income of $35k should expect to earn $20k after conviction, due to the effect that this conviction has on his reputation. A similar phenomenon has been found to apply to firms: the negative reputational effects that firms endure after being accused or convicted of fraud tend to cause a substantial loss in

stock value (Karpoff & Lott, 1993). Such findings suggest not only that information indicative of untrustworthiness tends to have a large effect on one's reputation, but also that it may be quite costly to be thought of as untrustworthy. I return to this in the next chapter.

By relying on third parties, observers may become aware of signs of a person's trustworthiness that they did not notice themselves. Most importantly, they may learn about that person's past choices in self-benefiting opportunities. But even when third parties have identical information, observers may increase their accuracy by relying on their judgments. Given that people have some accuracy in judging the trustworthiness of others, observers will usually increase their accuracy by incorporating the judgments of others. As Spiekermann (2007) shows by means of an application of Condorcet's jury theorem, even when the probability that a person is correctly recognised by an individual observer is only slightly above chance performance, he is likely to be recognised if a group of observers aggregate their judgments. Taking this together, I conclude that, in comparison with isolated interactions, we may expect persons to be more translucent when their interaction partners have access to third party judgments.

3.3 Evaluating the Translucency Assumption for informed interaction

I have argued that persons are more translucent in informed interactions than they are in isolated interactions: when a person's observers are informed by past observations or by third parties, they are more likely to judge her disposition correctly. For constrained maximizers this means that they are unlikely to be mistaken for being untrustworthy in informed interactions. Findings in the previous section suggested that in isolated interaction with explicit communication, trustworthy persons are recognised about 85% of the time. The above considerations provide reason to think this percentage is even higher for informed interaction.

There is reason to think that a straightforward maximizer's translucency will increase more in informed interactions than a constrained maximizer's. I assume that a straightforward maximizer cannot prevent his interaction partners sometimes learning about his past choices with respect to self-benefiting opportunities.[14] These persons may either have observed the

[14] This is not to assume that a straightforward maximizer does not take into account possible negative consequences of violating in individual cases. Indeed, it is consistent with my analysis that a straightforward maximizer will not violate in an interaction when he expects that the negative consequences for his future opportunities are higher than the direct benefits of

violations themselves, or heard about them from others. As the straightforward maximizer's number of violations increases with the passing of time, so does the likelihood of interaction partners learning about it. The trust-asymmetry principle suggests that once an interaction partner has this information, he will judge the straightforward maximizer to be untrustworthy. But even without knowledge of past transgressions the straightforward maximizers's interaction partners are more likely to recognise his untrustworthiness in informed than isolated interactions, because they are more likely to have observed signs of his untrustworthiness themselves, learned about them from third parties, or both.

Two other facts about trust worsen the straightforward maximizer's plight further. First, lost trust is not easily regained (Dasgupta, 2000; Slovic, 1999). Trust typically rebuilds slowly and often does not return to its original level. While there is reason to think that with time, amends and consistently trustworthy choices, a degree of trust may return (Schweitzer et al., 2006), this is of little help to the straightforward maximizer since trustworthy choices are not in his repertoire. Second, people do not only forgive easily, they also do not forget easily. Individuals confronted with pictures of people marked as either 'cheater' or 'co-operator' tend to remember pictures of cheaters better than those of co-operators (Oda, 1997; Mealey, Daood, & Krage, 1996). Interestingly, this memory effect has even been found when pictures of individuals who cooperated or cheated were *not* explicitly marked as such (Verplaetse et al., 2007; Yamagishi, Tanida, Mashima, Shimoma, & Kanazawa, 2003). This suggests that we have an enhanced ability not just to remember persons who have been shown to be untrustworthy, but also for persons who *look* untrustworthy.

The crucial question now is whether this all means that the Translucency Assumption is plausible for informed interaction. I rely again on Gauthier's model. As was mentioned before, for constrained maximization to top

---

defecting. Such an expectation affects the payoffs, and implies that the interaction does not involve a self-benefiting opportunity. The reason why a straightforward maximizer's translucency can be expected to increase in a non-isolated setting has to do with information that agents possess in this new setting. That is, to move away from isolated interaction to informed interactions introduces additional uncertainties. In particular, I assume that agents when facing self-benefiting opportunities may not always know that violating would affect the judgments of future interaction partners of his. His knowledge about the implications that a given violation has on the judgments of future interaction partners is limited. For example, he may not be aware that a person whom he exploits at one point in time is someone that he will face again at a later point in time. I discuss the question of how much information we must assume agents to have more extensively in the next chapter (§8.3.1).

straightforward maximization in a symmetric PD (see Table 2 in §2.5) the probability $\sqrt{p}$ that a constrained maximizer is recognised by constrained maximizers must be at least twice as high as the probability q that a straightforward maximizer is not detected by constrained maximizers. For values of r below 1, this difference must be larger.[15] Do the above findings support thinking that $\sqrt{p}$ is more than twice as high than q in non-isolated interaction?

To be sure, I have in this section not discussed studies that provide direct support for such estimates. But the above conclusions do provide support for such estimates indirectly. In the previous section I derived estimates for $\sqrt{p}$ and q of respectively 0.78 and 0.54 for isolated interaction. The conclusions presented in this section provide reason to think that these values change substantially in informed interaction. First, the value of $\sqrt{p}$ should be expected to be substantially higher than in isolated interactions. With additional information about her disposition available to interaction partners, it should become rather unlikely that a constrained maximizer is mistaken for a straightforward maximizer by other constrained maximizers. Second, the value of q should be expected to be substantially lower than in isolated interaction. Given that untrustworthy persons are much more likely to be recognised as such when their interaction partners have observed them previously and/or have access to the judgments of others, we may expect that in informed interaction the probability q that straightforward maximizers are not detected as such by constrained maximizers decreases substantially below that associated with isolated interaction. This means that with respect to informed interactions, it is much more plausible that constrained maximization is more advantageous than straightforward maximization than it is with respect to isolated interactions. For example, if the probability $\sqrt{p}$ increases from 0.78 in isolated interaction to a value of 0.9 or higher in informed interaction, and the probability q decreases from 0.54 to a value of 0.4 or lower, constrained maximization is more advantageous than straightforward maximization provided only 1 out of 5 interaction partners is untrustworthy ($r = 0.8$).[16]

---

[15] It can be calculated that were we to take r to be $2/3$, $\sqrt{p}$ needs to be 2-and-a-half times as high as q (see the 'Gauthier condition' in the appendix, §3). If on the other hand we take r to be $3/4$, $\sqrt{p}$ needs to be 2-and-a-third times as high as $\sqrt{p}$.

[16] This is on the assumption that interactions have the structure of the symmetric PD (see 6§4, or §3 of the appendix).

A concern may be raised at this point. The above argument is based on evidence and considerations that apply to people in general. Even if the above values are correct as interpersonal averages, it does not exclude there being a minority for whom the probability of being detected as a straightforward maximizer by constrained maximizers is sufficiently low that straightforward maximization is more advantageous for them. Put differently, there may be a minority for whom the probability q is above 0.4. As I mentioned in the previous chapter, Sayre-McCord (1991) has suggested there in fact is such a minority of 'deceptive people'. Furthermore, there is some empirical support for individual differences in translucency. As I mentioned when discussing lie detection, some persons are substantially less likely to be detected when lying than people on average (Bond & DePaulo, 2008). May such persons be sufficiently skilled deceivers to keep their q above 0.4? As the next section will introduce considerations relevant to this issue, I return to this there.

For now I conclude that, provided interactions are informed interactions with constrained maximizers that have the structure of a symmetric PD, *most* persons may, due to translucency, expect to be better off as constrained than as straightforward maximizers. Clearly, the extent to which this conclusion supports the Translucency Assumption depends on whether or not (1) interactions are typically informed and (2) agents interaction partners are typically constrained maximizers. I shall now discuss the plausibility of these conditions.

4 Interaction control and trust

The crucial thought behind Gauthier's claim that constrained maximization is advantageous is that it enhances a person's opportunities for cooperation. As Gauthier writes in a text published after *Morals by Agreement*, "[i]n his interactions, others are lead, through knowledge or belief about his dispositions, intentions, and plans, to behave in ways that enable him to do better" (1991b, p. 328). As we have seen, in his formal analysis Gauthier incorporates this point in terms of the likelihood that a person's interaction partners will cooperate: a constrained maximizer's interaction partners are more likely to cooperate than a straightforward maximizer's. This is, however, just one of several ways in which one's disposition may affect one's opportunities. In the present section I shall show that Gauthier's argument can be strengthened if we take into account additional effects.

One such effect is already mentioned by Hobbes. As we saw in Chapter 6 (§2), Hobbes responds to the Foole that:

> He [...] that breaketh his Covenant, and consequently declareth that the thinks he may with reason do so, cannot be received into any Society, that unite themselves for Peace and Defence, but by the errour of them that receive him; nor when he is received, be retained in it, without seeing the danger of their errour; (Hobbes, 1651/1991, p. 102)

Hobbes warns the Foole of the risk of social ostracism. When a person who does not keep his Covenant is detected violating it, he will be kept out of civil society. Hobbes seems to mean not just that others are less likely to respond cooperatively to such a person; he means that such a person will be excluded from interacting with them altogether.

Hobbes's response points to a fact that is not explicitly accounted for in Gauthier's analysis: that people can choose with whom to interact. They can collectively exclude untrustworthy persons because they can individually choose to refrain from interacting with them. The point can be made more general, however. We do not simply find ourselves in interactions in which we can either choose to cooperate or to defect; we have *control* over the nature of our interactions, including with whom we have them.

I shall show in this section that Gauthier's argument for the rationality of constrained maximization can be made more plausible if we include the assumption that persons have control over their interactions. More precisely, I shall argue that if we combine this assumption of interaction control with the findings on translucency, the two conditions with which the previous section ended are met. In the second subsection I shall discuss what interaction control means for straightforward maximization. I shall also return there to the aforementioned concern regarding persons skilled in deception.

4.1 Constrained maximization and interaction control

The starting point of this section is that self-benefiting opportunities do not arise from thin air. They result from choices made by each of the parties involved. This is clearly the case for self-benefiting opportunities that come into being because of explicit agreements. By making an agreement, parties create a situation in which each of them can choose to comply with or violate the agreement. This happens when parties sign a non-binding contract, but also when promises are made. Self-benefiting opportunities may also result from implicit agreements. Such an agreement is made, for example, when one

party performs a service for another party on the shared understanding that it will be reciprocated at some later time.

Self-benefiting opportunities may also arise without the parties involved having engaged in an agreement. In the context of a community governed by public norms of cooperation, an individual may expect another individual to observe these norms even though he has not made an explicit agreement with him about this. For example, when I invite someone to my house, I do this in the expectation that he will not harm me or take my property. While I could not avoid being in this situation by refraining from making some agreement, I could avoid it by not providing him access to my house in the first place. Indeed, I may have been able to avoid interacting with him in the first place.

The point is that we may assume persons to have control over whom they give self-benefiting opportunities to. Persons have control over when and where to be vulnerable for exploitation. They can refrain from engaging in agreements with others when they do not want to. They can, at least to a certain degree, avoid environments in which they expect they may be forced into interactions they do not want to be in. And when they happen to find themselves in an environment in which such interactions may occur, they may yet reduce the risk by keeping others at a social or a physical distance.

Combined with the idea of translucency, the assumption of interaction control is sufficient to justify thinking that the two conditions that must be satisfied for constrained maximization to trump straightforward maximization are in fact satisfied. Consider first the condition that most of one's interactions are informed. The above considerations imply that one can *ensure* that most of one's interactions are informed, at least if one chooses to be a constrained maximizer. On the one hand, a constrained maximizer may expect to be generally able to avoid giving self-benefiting opportunities to unfamiliar parties. She may avoid isolated interactions. On the other hand, she may expect to have little difficulty in gaining access to informed interactions. Especially when others know about her history, they are likely to recognise her trustworthiness.

Consider now the condition that most of one's interaction partners are constrained maximizers. Just as persons can refrain from having isolated interactions, they can refrain from having informed interactions with individuals they do not trust. That is what Hobbes presumes when he warns of the risk of ostracism. But persons do not have to wait avoiding another until he has actually committed a violation. Gauthier's insight, confirmed by empirical findings, is that persons may detect untrustworthy others before

being exploited by them. Combining Hobbes and Gauthier, we realise that persons may to a large extent avoid interacting with untrustworthy persons by interacting solely with persons who look trustworthy. Indeed, they may go as far as to avoid interacting with anyone appearing less than fully trustworthy in the light of information from previous observations or third parties. This approach should be quite effective in avoiding straightforward maximizers. Due to their translucency, straightforward maximizers tend to look less trustworthy than constrained maximizers. By adopting the cautious approach, straightforward maximizers only have to spark the slightest doubt regarding their trustworthiness to be avoided as interaction partners.

The cautious approach will of course also increase the probability of avoiding interacting with persons who are in fact trustworthy. This may involve costs. When I find myself in need of using the washroom while travelling by train, it may lead me, cumbersomely, to take my luggage with me into the washroom. Persons can minimise the occurrence of such situations, however, by surrounding themselves with persons they trust. Given translucency such persons are also likely to be trustworthy. Of particular importance in this regard is the possibility of developing personal relations. Trust lies at the basis of our personal relations with acquaintances, colleagues, neighbours, and friends. Persons who have such relations with one another stand in what we may call a trust relation with one another. Through these relationships we can also come to know new parties with whom we can develop trust relations. Persons whom we trust will have trust relations with third parties with whom we do not yet stand in such a relation but whom we have reason to trust, provided that those we trust have good judgment. Indeed, persons who trust one another tend be become part of networks of interconnected trust relations. By becoming part of such a 'trust community', persons can become part of a community that contains a high proportion of trustworthy individuals (cf. Spiekermann, 2007).

I concluded in the previous section that if one's interactions are typically informed rather than isolated and are typically with constrained maximizers rather than with straightforward maximizers, persons are sufficiently translucent for constrained maximization to trump straightforward maximization. The above shows that, on the assumption of interaction control, we may expect these conditions to be met; or at least, they may be expected to be satisfied for constrained maximizers. It may be thought, however, that the introduction of interaction control also provides new opportunities for the straightforward maximizer.

4.2 Straightforward maximization and interaction control

Interaction control has first and foremost a negative consequence for the straightforward maximizer. It is not hard to see that if interaction control enables constrained maximizers to increase the probability of interacting with fellow constrained maximizers and to decrease that of interacting with straightforward maximizers, the straightforward maximizer's chance of interacting with a constrained maximizer decreases. Indeed, given that straightforward maximizers tend to look less trustworthy than constrained maximizers, they are less likely to be welcomed as interaction partners by constrained maximizers. This effect should be particularly large with regard to interactions that take place in an informed setting. Not only are potential interaction partners more likely to have detected signs of their untrustworthiness, they will also have received little evidence of the opposite; straightforward maximizers do not perform trustworthy actions. This means also that straightforward maximizers are unlikely to develop trust relationships and gain access, and be retained, in trust communities. Straightforward maximizers are therefore less likely to interact with constrained maximizers than constrained maximizers.

This point can easily be accounted for in Gauthier's analysis on which I have relied in the previous section. Instead of using the same value for the probability r of one's interaction partner being a constrained maximizer when calculating the expected utility of the two dispositions, the implication of the above consideration is that the expected utility of straightforward maximization should be calculated with a lower value of r than that of constrained maximization. The effect of this change is that, everything else being equal, the expected advantage of straightforward maximization decreases. This also implies that the degree of translucency required for constrained maximization to top straightforward maximization decreases (see the appendix, §4). In particular, the probability $\sqrt{p}$ that a constrained maximizer is recognised now no longer needs to be twice as high as the probability q that a straightforward maximizer remains undetected.

I take the above point also to go a long way in resolving the concern that persons with exceptional deception skills are better off as straightforward maximizers. The crucial idea is that a straightforward maximizer does not need to be detected as such for his disposition to negatively affect his opportunities. Straightforward maximizers must not only deceive constrained maximizers into cooperating to get the benefits of exploitation; they must first

persuade them to interact at all. Even persons with exceptional deception skills may expect to experience more difficulty here if they are straightforward maximizers. Although they are less likely to be detected than others, even skilled liars are frequently detected lying (Bond & DePaulo, 2008).

More importantly than that, refraining from looking untrustworthy by avoiding detection is quite another thing than looking trustworthy. As straightforward maximizers do not perform trustworthy actions, the only way in which a deceptive person may develop a reputation of being trustworthy is through deception and trickery. As he will not live up to this reputation by actually cooperating, he is unlikely to develop trust relations. Given these considerations, Sayre-McCord's (1991) contention that deceptive persons tend to "travel with a glowing reputation" is not very plausible (p. 192). In contrast, a deceptive person should expect to look less trustworthy than constrained maximizers do were he to settle for straightforward maximization. In that case, he is less likely to be interacting with constrained maximizers.

There is another way in which the interactions that one would have as a straightforward maximizer may be expected to differ from those one would have as a constrained maximizer. Trust does not only affect whether persons are willing to take a risk with others or not, but also how large a risk they are willing to accept. When I need to use the washroom when travelling by train, my decision to leave my luggage with a fellow traveller whom I do not know will not only be based on my impression of him but also on what I am carrying in my suitcase. If it is dirty laundry, I am likely to take the risk of being exploited. But were carrying my MacBook, I would decide differently. It is well established that persons become increasingly risk averse when the stakes are higher (Holt & Laury, 2002). Unsurprisingly, what stakes persons are prepared to accept in an interaction has been found to depend on how much they trust their interaction partners (Johansson-Stenman, Mahmud, & Martinsson, 2005; Parks & Hulbert, 1995). Whereas they will have little problem in taking large risks with persons with whom they have developed a trust relation, they may be careful taking even small risks with persons they have little reason to trust. Given that constrained maximizers look more trustworthy and are more likely to have trust relations than straightforward maximizers, including those with excellent deception skills, the interactions of constrained maximizers will on average involve higher stakes.

In his formal analysis, Gauthier uses one identical payoff matrix to represent all interactions, independent of one's disposition. The implication of the above consideration is that we should use a different payoff matrix for

calculating the expected advantage of constrained maximization than for straightforward maximization. The effect of this change is, once again, that, independent of other factors, the expected advantage of constrained maximization increases with respect to that of straightforward maximization (see the appendix, §4).

I have argued that, in comparison with constrained maximizers, the interactions of straightforward maximizers involve lower average stakes and are less likely to include constrained maximizers. These points imply that constrained maximization is more advantageous than straightforward maximization *even* if straightforward maximizers have a relatively low degree of translucency. Indeed, if we increase the stakes of constrained maximizers' interactions by 25% relative to those of straightforward maximizers and furthermore assume that constrained maximizers interact with other constrained maximizers 90% of the time and straightforward maximizers 75% of the time, constrained maximization is more advantageous than straightforward maximization even if straightforward maximizers remain undetected in interactions 70% of the time (I show this in the appendix, §4). Put differently, even straightforward maximizers who are exceptionally skilled at deception are worse off than constrained maximizers.

It may be objected at this point that the expected advantage of being a straightforward maximizer is not lower than that of being a constrained maximizer if we take into account that interaction control also gives the straightforward maximizer new opportunities. In particular, a straightforward maximizer can choose to have mostly isolated rather than informed interactions, in which he may expect to be less translucent. The straightforward maximizer does not need to settle in a community, but can stay on the move instead, always on the lookout for opportunities to exploit others. As we saw in section 3, findings on translucency suggest that in isolated interaction straightforward maximization is more advantageous than constrained maximization. May persons not do better by being straightforward maximizers who have mostly isolated interactions?

The objection fails. It presupposes that the isolated interactions one would have as a straightforward maximizer are just as fruitful as the informed interactions one may have as a constrained maximizer. But this is implausible. Persons in isolated interactions have to form their judgments of each other on the basis of cheap talk and behavioural signs. Even when they judge one another as trustworthy, the level of trust will typically be low. Isolated interactions may therefore be expected to have on average lower stakes than

informed interactions. Furthermore, isolated interactions may be expected to involve a lower proportion of constrained maximizers than informed interactions. As I argued in the previous subsection, constrained maximizers can and will choose to have mostly informed interactions rather than isolated interactions. Straightforward maximizers and other social predators, on the other hand, may be attracted to isolated interactions due to them being less translucent in such a setting. This also means, however, that there is an increased risk for mistaking fellow straightforward maximizers for constrained maximizers whom one may seek to exploit.

In addition, there is reason to expect that isolated interactions with self-benefiting opportunities may not come by very frequently. While persons who trust one another see little problem in giving each other self-benefiting opportunities, this is very different when there is little or no trust. People are careful when facing others they do not know. As Hobbes writes:

> Let him therefore consider with himselfe, when taking a journey, he armes himselfe, and seeks to go well accompanied; when going to sleep, he locks his dores; when even in his house he locks his chests. (Hobbes, 1651/1991 p. 89)

People tend to protect themselves against being exploited. They tend not to give self-benefiting opportunities to others they do not trust.[17] In public contexts, for example, they tend to keep both a physical and a social distance from strangers so as not to make themselves vulnerable to exploitation. For these reasons, when a straightforward maximizer faces a potential victim he will often not be able to get an opportunity to exploit that person *even* if he remains undetected as being untrustworthy. While this point may not be easily incorporated into Gauthier's formal analysis, a lower number of self-benefiting opportunities clearly implies a decrease in the expected advantage of straightforward maximization in comparison with that of constrained maximization.

I conclude that the assumption of interaction control does not benefit straightforward maximizers to the same extent as it benefits constrained

[17] Against this it may be objected that there are persons who are much easier to fool into interacting. A straightforward maximizer may concentrate on exploiting such easy victims. The question is whether the benefits associated with such interactions are higher than the cooperative benefits one may have as a constrained maximizer. This depends among other things on the number of interactions one may expect to have with easy victims. My response to this objection is similar to the one put forward in this paragraph. Although it is an empirical question fully answering which requires additional data, I expect that there are relatively few easy victims around, and that exploiting them would not compensate for the straightforward maximizer's loss in cooperative opportunities.

maximizers. In fact, it reduces the expected advantage of their disposition as it gives potential victims the power to avoid interacting with them and to keep the stakes of interactions low. Even persons with exceptional skills of deception may therefore expect to be better off as constrained than as straightforward maximizers.

5 Conclusions

This chapter examined two challenges that have been posed against the Translucency Assumption. The first challenge, that the idea of translucency is psychologically implausible, was quickly defeated: empirical studies show that persons have surprising skill in predicting whether others are trustworthy or not, even if they are strangers to each other. The second challenge, that persons are insufficiently translucent for constrained maximization to be more advantageous than straightforward maximization, proved to be weightier. Although studies show strangers to be moderately translucent, this is less than Gauthier's argument requires. The upshot is that constrained maximization is not advantageous if one is to have only isolated interactions.

Matters change, however, if we add some realism to the analysis. I first argued that we should take into account that most of a person's interactions are not isolated encounters with strangers. Translucency increases in what I called informed interactions: when interaction partners have some familiarity with each other or when they can take judgments of third parties into account. I argued that it is plausible that this degree of translucency is such that, if a person's interactions are usually informed and if his interaction partners are usually trustworthy, being a constrained maximizer is more advantageous than being a straightforward maximizer.

I then argued that once we assume persons to have control over whom they give self-benefiting opportunities to, these two conditions are satisfied. With interaction control, the constrained maximizer can ensure that the majority of her interactions are informed and with other constrained maximizers. In addition, for the straightforward maximizer interaction control introduces a risk of being ostracised, further reducing his expected utility. I thus conclude that once the idea that interactions are typically isolated is abandoned, Gauthier's claim that our translucency makes being a constrained maximizer more advantageous than straightforward maximization empirically plausible. The second challenge to the Translucency Assumption can thus be disposed of.

8

Why Not Be an Opportunist?

1 Introduction

Several critics of Gauthier have pointed out that even if it is more advantageous to be a constrained maximizer than a straightforward maximizer, there may be alternative dispositions that are more advantageous than constrained maximization. One plausible candidate of such a disposition is what David Copp calls a reserved maximizer:

> A reserved maximizer has exactly the disposition of a constrained maximizer, except that he will violate a requirement of a cooperative scheme whenever he has the opportunity to win the jackpot. He will take opportunities to make very great gains in utility, when the probability of detection is very low. For example, unlike a constrained maximizer, he may steal the money from a lost wallet, provided enough money is involved and provided he is quite sure he was not observed finding the wallet. A person might do better as a reserved maximizer than as a constrained maximizer. (Copp, 1991, p. 221)

Like the constrained maximizer, and unlike the straightforward maximizer, the reserved maximizer is conditionally cooperative. Unlike the constrained maximizer, his cooperation is not only conditional on the disposition of his interaction partners but also on what there is to gain by violating. If the difference in advantage between complying and violating is large and the probability of the violation being detected is very low—if he has what is often called a *golden opportunity*—he will violate rather than comply.[1]

It is not hard to see why Copp thinks that reserved maximization might be more advantageous than constrained maximization. Constrained and reserved maximizers only choose differently with respect to those self-benefiting opportunities that constitute golden opportunities. As reserved

[1] Robert Frank (1988) uses the term in this way, as do Gregory Kavka (1995) and Christopher Morris (1999).

maximizers defect in such situations and constrained maximizers cooperate, reserved maximizers do better with respect to the latter *unless* the others involved recognise reserved maximizers first and prevent being exploited by them. However, given that reserved maximizers choose in other self-benefiting opportunities like constrained maximizers, it seems that plausible they can develop a reputation that makes others think they are trustworthy. If this is indeed so, being a reserved maximizer with regard to golden opportunities is more advantageous than being a constrained maximizer. In that case, it seems that reserved maximization must also be a more advantageous disposition overall.

Suppose reserved maximization does indeed trump constrained maximization, what would that mean for Gauthier's argument? This depends on whether we interpret reserved maximization to be an amoral disposition or not. On one interpretation, reserved maximization is what Gauthier calls "straightforward maximization in its most effective disguise" (p. 169). It is the disposition of "the person who, taking a larger view than her fellows, serves her overall interest by sacrificing the immediate benefits of [...] violating co-operative arrangements in order to obtain the long-run benefits of being trusted by others" (p. 169). Gauthier insists that such a person is not moral or, as he puts it, "such a person exhibits no real constraint" (pp. 169-170).[2] Contrary to the constrained maximizer, she has not internalised moral principles as constraints on the pursuit of her interest; she only acts in conformity with moral norms because it suits her interests.

If this amoral disposition turns out to be more advantageous than constrained maximization, Gauthier's argument for the rationality of morality fails. It would mean that merely *appearing* moral is more advantageous than actually *being* moral.

There is, however, an alternative interpretation of the description described by Copp which does not yield this conclusion. Gauthier describes the constrained maximizer as taking himself to have *decisive* reason to constrain the pursuit of his self-interest, provided others do so as well. Copp may be interpreted as describing a person who, like the constrained maximizer, takes himself to have good reason to comply with cooperative norms, but only up to a point: when violating is very advantageous, this may overrule the reasons for cooperating. Although this agent may be described as less moral than the constrained maximizer, it surely goes too far to call him amoral. I will call this

[2] Gauthier makes these remarks when explaining how we should *not* interpret constrained maximization.

disposition *semi-constrained maximization*; I shall use reserved maximization to refer to the amoral disposition described earlier.

In making the distinction between reserved maximization and semi-constrained maximization, I have added an element to the notion of disposition to choose that did not play a role in the previous two chapters: namely, the reasoning that underlies choices.[3] The constrained, semi-constrained, and reserved maximizer reason differently with respect to self-benefiting opportunities, even if they often choose identically. The constrained maximizer, Gauthier writes, "has internalized the idea of mutual benefit, so that in choosing his course of action he gives primary consideration to the prospect of realizing the co-operative outcome" (p. 157). He takes himself to have decisive reason not to take advantage of the cooperation of others, even if doing so is very advantageous for himself. The reserved maximizer, on the other hand, has not committed to or internalised moral principles: she cooperates only because it is the best way to protect her long-term interests. Consequently, she sees *no reason* not to exploit others when it will not have negative consequences for herself, such as when she has a golden opportunity. Finally, like the constrained maximizer, the semi-constrained maximizer takes himself to have reason to cooperate and not take advantage of others. However, for him this is not a decisive but a *pro tanto* reason, that may on occasion be trumped by his individual advantage. Semi-constrained maximization thus only challenges the rationality of always giving moral considerations *priority* over self-interest. Although both alternative dispositions challenge Gauthier's argument for constrained maximization, reserved maximization therefore presents the more serious threat to Gauthier's defence of the rationality of morality.

My investigation concentrates on reserved maximization. I take the argument for the rationality of reserved maximization to have two crucial premises. The *first* premise is that reserved maximization is substantially more advantageous with regard to situations that involve golden opportunities. The argument as presented above supposes furthermore that reserved maximizers do about as well as constrained maximizers with regard to other interactions. However, this is not required for reserved maximization to be more advantageous: the argument only requires that the gain with respect to golden opportunities is larger than losses in cooperative opportunities. This is the *second* premise of the argument.

[3] This element does, however, play an important role in Gauthier's analysis. Gauthier insists the constrained maximizer "reasons in a different way" from the straightforward maximizer (p. 170).

The plausibility of this second premise is the prime target of my investigation. Just as with straightforward maximization, I shall take reserved maximization to be associated with a type of untrustworthiness: given that the reserved maximizer will violate whenever he believes he has a golden opportunity, he cannot be trusted to cooperate in the way that a constrained maximizer can be trusted. The second premise requires that others would either happily cooperate with such a person, or that they are not very good at recognising this type of untrustworthiness. I shall assume that the first condition is not met; that persons in general are hesitant to cooperate with reserved maximizers, and prefer to cooperate with constrained maximizers instead. This assumption is plausible for two reasons. First, cooperating with a reserved maximizer involves a risk, as he will defect whenever he believes he has a golden opportunity. Persons thus have reason to be hesitant to cooperate with reserved maximizers, especially if they can have equally fruitful interactions that do not involve such risks with constrained maximizers. Second, there is reason to think persons in fact are hesitant to cooperate with others whom they believe to be reserved maximizers. Not only is there extensive evidence that persons prefer not to cooperate with untrustworthy others, persons appear to also hold very negative attitudes towards the sort of opportunism associated with reserved maximization (Tetlock, 2000; Tetlock 2003). Whether the second condition is met depends on whether reserved maximizers are translucent. In the light of the findings presented in the previous chapter, it is far from obvious they are not.

I will discuss the translucency of reserved maximizers in the following two sections. The next section considers whether reserved maximizers may be distinguished from constrained maximizers without having been detected violating. The third and the fourth section discuss to what extent the reserved maximizer's translucency should be expected to increase due to his propensity to violate whenever he believes he has a golden opportunity. In the fifth section I will discuss the implications of my findings for the challenge posed by reserved maximization. Finally, section six returns to the issue of individual differences, and discusses whether there are persons who may expect to do considerably better as reserved maximizers than others.

Before starting, it should be noted that in this chapter I will not make use of a formal analysis. Instead, I will give a non-formal assessment of the argument for reserved maximization. I will draw on a distinction between interactions that do and interactions that do not involve golden opportunities,

and examine whether the expected advantage of reserved maximization with regard to the latter kind makes up for potential losses in the former kind.

2 The signs of reserved maximization

The second premise of the argument for reserved maximization requires that reserved maximizers look similar to constrained maximizers. I will investigate whether this is the case, starting with the question of whether reserved maximizers who have not been yet been detected violating look just as trustworthy as constrained maximizers or not.

The previous chapter reviewed empirical studies showing that, even without prior interactions or familiarity, people can with moderate accuracy predict whether they can trust others to cooperate or not (Brosig, 2002; i.e. Frank et al., 1993; Verplaetse et al., 2007). Communicating with them, or even merely observing their nonverbal behaviour, is sufficient for persons to have a sense of whether others are trustworthy or not. This supports the idea that untrustworthy and trustworthy persons can be distinguished due to signs associated with these dispositions. The question for now is whether there is reason to think that this applies also to the particular type of untrustworthiness associated with reserved maximization.

Let me start with an argument for why the reserved maximizer may *not* look different from a constrained maximizer in the situations we are interested in. The empirical studies just mentioned show that persons have some ability to distinguish between persons who choose to cooperate and persons who choose to defect. This is likely to do with the fact that such persons have different intentions. In the situations we are presently considering, reserved maximizers and constrained maximizers have identical intentions, however: they both intend to cooperate. These studies do therefore provide no reason for thinking that reserved maximizers and constrained maximizers look different as long as the former has not been detected violating.

In response to this, it should be pointed out that intentions are unlikely to be the only source of signs associated with a person's disposition. Studies in the 'thin slices' paradigm show that people have some ability to identify dispositions related to trustworthiness and untrustworthiness on the basis of nonverbal cues not directly associated with cooperating or violating (7§2.3). Several studies have found that we have some accuracy in distinguishing altruists from non-altruists on the basis of short videos that were taken in a context that did not involve such behaviours (Fetchenhauer et al., 2010; Oda et

al., 2009). Other studies have found that we are sensitive for the so-called dark triad personality types of narcissism, machiavellianism, and psychopathy (Back et al., 2010; Fowler et al., 2009; Holtzman, 2011). Given that they are in part characterised by a disposition to deceive and to manipulate and by a disregard for morality, these dispositions show a resemblance to reserved maximization.

Moreover, there are differences between the dispositions of reserved and constrained maximization that may be detected by observers. First of all, as I explained in the introduction, reserved and constrained maximizers reason differently with respect to self-benefiting opportunities. The constrained maximizer takes herself to have decisive reason not to exploit others, provided others may be expected to cooperate. By having internalised moral principles that govern her choices, the constrained maximizer will typically not even contemplate violating when facing a self-benefiting opportunity. The reserved maximizer, on the other hand, remains fully open to the possibility of violating when facing a self-benefiting opportunity. Indeed, she must *calculate* the expected benefit of defecting and the probability of being detected in order to determine whether it is a golden opportunity or not.

People associate calculating with untrustworthiness. As Bennis and colleagues (2010) write in a review, "even knowing that a third party merely contemplated [a trade off of sacred values and money] elicits contempt, disgust, and a desire to punish from participants" (p. 190). This seems to be especially the case when we had placed trust in the people who do so. As Randolph Nesse (2001) writes:

> Perhaps the strongest evidence that friendships are based on commitment and not reciprocity is the revulsion people feel on discovering that a friend is calculating the benefits of acting in one way or another. People intuitively recognize that such calculators are not friends at all, but exchangers of favors at best, and devious exploiters at worst. (Nesse, 2001, p. 31)

While there are to my knowledge no studies regarding our ability to distinguish truly trustworthy persons from persons who cooperate out of calculation, it is likely that there are observable differences between such persons. A calculating reserved maximizer needs to consider aspects of his situation, such as the benefit of violating and the probability of being detected, that are irrelevant to constrained maximizers. The process of considering such aspects may involve visible components. For example, before deciding whether to return a found wallet, a reserved maximizer will look at how much money is in it, which is usually not relevant for a constrained maximizer. Furthermore, considering these aspects will require time and cognitive resources. While it

may often be instantly clear for the reserved maximizer whether he does or does not have a golden opportunity, on other occasions it will require him to engage in an effortful process of deliberation and calculation that a constrained maximizer would never engage in.

A second difference that may be picked up by observers concerns the emotional tendencies of reserved and constrained maximizers. As I explained in the previous chapter (§2.5), by relating constrained maximization to trustworthiness I take the former, like the latter, to include an emotional component. With Gauthier, I assume that constrained maximizers are "affectively engaged by compliance, so that the familiar feelings of respect and resentment, of self-respect and guilt, are linked appropriately with the fair and unfair behavior of others and oneself" (1986, p. 266). As reserved maximizers do not have internalised moral principles nor take themselves to have reason to be moral, they will not have such emotions.

Besides this difference in moral sentiments, constrained maximizers and reserved maximizers may be expected to differ in the extent to which they develop what Robert Frank (1988; 2005) calls 'sympathetic bonds'. As Frank points out, sympathy increases one's motivation to cooperate:

> A person who is sympathetic toward potential trading partners is, by virtue of that concern, less likely than others to yield to temptation in the current interaction. Such a person would still find the gains from defecting attractive, but their allure would be mitigated by the prospect of the immediate aversive psychological reaction that would be triggered by defecting. (Frank, 2005, pp. 92-93)

For constrained maximizers, the phenomenon of sympathetic bonding will make it easier to do what they believe they have reason to do.[4] For reserved maximizers, in contrast, it poses a difficulty. When a reserved maximizer would come to identify with another person's interests, he may experience difficulty taking advantage of her when the opportunity arises. Reserved maximizers should therefore avoid, or at least be reserved, in developing sympathetic bonds with others.

Observers may recognise such emotional differences, we saw in the previous chapter, and take them into account in their judgments of trustworthiness (§2). A person's expressions of moral sentiments and of sympathy appear to play an important role in how much he is trusted (Frank, 1988; Ross & Dumouchel, 2004). As persons have only limited controllability

[4] As I noted before, I assume that whether a person has a self-benefiting opportunity depends only on her self-regarding interests, and not on any prosocial or moral sentiments she may have.

of emotional expressions, they will sometimes inadvertently reveal their true feelings. Empirical studies show that observers are surprisingly good at noticing such nonverbal information, and that they rely on it when judging trustworthiness. We thus have reason to expect that due to these emotional differences reserved maximizers look less trustworthy than constrained maximizers.

A third difference between reserved maximizers and constrained maximizers that may affect the judgments of observers are additional actions that the former must perform before or after taking a golden opportunity. In the course of taking a golden opportunity, a reserved maximizer may need to rely on manipulation and deception. In some cases, a reserved maximizer will know while interacting with others that he will in the future have a golden opportunity to exploit them. Given this knowledge, he will have to act sincerely, even though he already has the intention to exploit the other. Similarly, after having taken a golden opportunity he has to make sure his violation remains undetected. For example, a reserved maximizer who has taken a substantial amount of money from a found wallet has to make sure others do not become suspicious about his increase in wealth. The findings from the previous chapter suggest that even if a reserved maximizer's taking of a golden opportunity remains undetected, such actions may spark distrust.

To conclude, there are several differences between reserved maximizers and constrained maximizers that should be expected to have the consequence that reserved maximizers look, generally speaking, less trustworthy than constrained maximizers. Even when they have not been detected violating, others are less likely to trust them than they are to trust constrained maximizers. Being a reserved maximizer may thus be expected to have a negative effect on a person's cooperative opportunities, even if he is not actually detected taking golden opportunities.

This of course does not imply that the argument for reserved maximization fails. While we should expect reserved maximizers to do less well than constrained maximizers in interactions that do not involve golden opportunities, the above considerations do not show how much less well they do. It may therefore still be the case that the benefits associated with taking golden opportunities compensate for these costs in cooperative opportunities. The following two sections discuss whether this should be expected to be so.

Importantly, not all of the above considerations apply in the same way to semi-constrained maximization, the other alternative disposition introduced above. In particular, given that semi-constrained maximizers take themselves

to have reason to act morally, they may have similar moral sentiments as the constrained maximizer. This may be taken to suggest that semi-constrained maximizers will also look more trustworthy than reserved maximizers. However, his moral sentiments may also work against the semi-constrained maximizer. Due to moral sentiments such as anticipatory guilt, it may be more difficult to deceive and manipulate others. Moreover, semi-constrained maximizers may occasionally experience actual guilt over undetected golden opportunities, in particular when they face victims of their actions. Only if semi-constrained maximizers are able to control their moral sentiments with respect to golden opportunities will they look more trustworthy than reserved maximizers.

3 The risk of detection

The most obvious way in which a reserved maximizer may lose the trust of others is when he is detected violating. The argument for reserved maximization supposes that, given that a reserved maximizer defects only when he believes the probability of detection to be very low, the probability of being detected is also very low. There are reasons for thinking that violating when and only when one has a golden opportunity is, however, not as easy as it may sound.

3.1 Apparent golden opportunities

The reserved maximizer can be thought of as combining the strengths of constrained and straightforward maximization. Unlike the straightforward maximizer, he recognises that one can only look trustworthy if one tends to choose cooperatively. But he also recognises that internalising moral principles prevents him from taking golden opportunities. He therefore does not internalise morality as a set of constraints, but sees it as a set of prudential rules with exceptions. Golden opportunities are self-benefiting opportunities that satisfy two additional conditions. The first is that violating is not just more advantageous than complying, but much more advantageous. In terms of the Prisoner's Dilemma, these are PDs in which there is a large difference between the payoffs of exploitation and mutual cooperation. The second condition is that the probability that the violation is detected is very low. More precisely, I take this condition to mean that the probability that the violation negatively affects future interactions is very low.

The expected advantage of reserved maximization depends on how often one would reap golden opportunities. This depends of course on how often such opportunities occur. However, it also depends on the reserved maximizer's ability to detect golden opportunities. Drawing on an argument made by Gregory Kavka (1995), I shall argue that detecting golden opportunities involves certain difficulties. While I concentrate on reserved maximization, the discussion below applies also to semi-constrained maximization.

The key idea is that in order to decide whether a self-benefiting opportunity is a golden opportunity, a reserved maximizer needs to form an assessment of the situation on the basis of available information. There is a similarity with constrained maximization. A constrained maximizer, Gauthier writes, "must estimate the likelihood that others involved in the prospective practice or interaction will act co-operatively, and calculate, not the utility she would expect were all to co-operate, but the utility she would expect if she co-operates, given her estimate of the degree to which others will co-operate" (p. 169). She makes this prediction on the basis of information she has about these others. The reserved maximizer must also make this prediction, but that is not all he must do. To assess whether the interaction involves a golden opportunity, he must determine the difference between the payoffs of exploitation and cooperation—in particular, he must determine whether it is above a certain threshold value. Furthermore, the reserved maximizer needs to determine whether the probability of detection is sufficiently low. He makes such predictions on the basis of information available to him.

To what extent is such information available? To answer this question, it is helpful to briefly say something about what the previous chapter's analysis assumed regarding the availability of information. Following Gauthier, I took agents to have complete information about direct payoffs of interactions. The reason for making this unrealistic assumption is simplicity. It is an innocuous assumption, as the argument in favour of constrained maximization does not appear to depend on it. By the same reasoning, it was *not* assumed that agents have complete information about the dispositions of their interaction partners. Transparency *does* favour constrained maximization in comparison with other dispositions. That is why Gauthier writes, "to assume transparency may seem to rob our argument of much of its interest… we shall have failed to show that under actual, or realistically possible, conditions, moral constraints are rational" (p. 174). Again by the same reasoning, and most relevantly here, I did not make the unrealistic assumption that agents have complete information

about the implications of their choices on their future interactions. One of the main reasons why straightforward maximizers are more translucent in informed interactions than in isolated interactions is that partners in such interactions may know about their past violations. But this may not occur were we to assume that straightforward maximizers always know whether violations will be detected or not. This assumption would thus make straightforward maximization appear more favourable than it would be under realistic conditions.

This last point applies just as well to reserved maximization. Complete information about whether a violation will be detected or not, or about how likely it is to be detected, would be extremely helpful for a reserved maximizer. Combined with complete knowledge of payoffs, the reserved maximizer would have a perfect ability for detecting golden opportunities. The assumption is, however, highly unrealistic. Persons typically only have limited information about the probability that a potential violation will be detected. This is partly due to their limited control over others. Persons have limited control over the extent to which others choose to observe and investigate them, and share their findings with others. Assuming complete information about the probability of violations being detected would thus, to paraphrase Gauthier, rob the argument in favour of reserved maximization of much of its interest; it would not show that under realistic conditions reserved maximization is more advantageous than constrained maximization. I shall therefore assume agents do not have complete information about the probability that a violation will be detected.

In effect, reserved maximizers sometimes have mistaken expectations about golden opportunities. They will sometimes expect self-benefiting opportunities to be golden opportunities even though they are not, and vice versa. Indeed, they may sometimes believe they have sufficient information to form such an expectation while in fact a crucial piece of information is missing to them. To use that helpful phrase of Rumsfeld again, there may be unknown unknowns that, had they been known, would have changed their expectations.[5] Christopher Morris gives a nice example of such a case:

> Consider next a story I once heard about some business practices in certain Asian markets where transactions were based on trust. Apparently, newcomers would be tested by offering them deals where they might be tempted to renege on their part of the arrangement and where they could, very easily, do so without expected loss. The

[5] See Chapter 5, footnote 3.

> purpose of the practice was to determine who might be a trustworthy business partner. Those who reneged would be excluded from future deals. (Morris, 1999, p. 93)

The point of this practice is to test the disposition of newcomers by giving them *apparent* golden opportunities. The practice works because individuals prone to taking such opportunities, such as reserved maximizers, tend neither to know they are being tested nor that they do not know this. Lacking this information, they mistakenly expect to have a golden opportunity and reveal their true colours.

As Kavka (1995) has pointed out, certain psychological tendencies that we have further increase the probability of having mistaken expectations about golden opportunities. Partly due to visceral responses, persons tend to overestimate the value of immediate rewards in comparison with later ones (e.g. Frank, 1988; Loewenstein, 2000). This is especially so when the latter are not certain but involve risk or uncertainty (e.g. Elster, 2007, pp. 114–115). Associated with this, there is in our minds a negative correlation between the benefits of activities and their risks: when we believe an activity to be beneficial, we tend to underestimate the risks it involves (Slovic & Peters, 2006). In addition, persons are prone to overestimate how well informed they are (Elster, 2007, p. 126). Finally, after having successfully reaped some golden opportunities, persons are likely to become overconfident. Due to these psychological tendencies, Kavka observes, persons are more likely to mistakenly think they have a golden opportunity than to pass up a golden opportunity out of fear of later detection.

The main point of this subsection is that, due to lack of information about the probability of violations being detected, persons cannot reliably identify golden opportunities. Furthermore, due to certain biases, they will over-detect golden opportunities. Consequently, reserved maximizers who violate whenever they expect to have a golden opportunity, would sometimes violate when the probability of being detected is in fact not low. As Kavka (1995) puts it, "in reality, many [...] apparent golden opportunities will turn out to be 'Foole's gold'" (p. 26).

3.2 Limits of self-control

I now turn to another imperfection of human agents that has a negative effect on the expected advantage of reserved maximization. It is not uncommon for persons to act contrary to what they think, or even have decided, to be best for them. Our ability to control ourselves in the face of temptations is limited.

There is reason to think that limited self-control makes it difficult for persons to only violate when they have a golden opportunity.

On the first page of *Morals by Agreement*, Gauthier approvingly cites Ogden Nash who writes, "O Duty, Why hast thou not the visage of a sweetie or a cutie?" When facing a self-benefiting opportunity, doing the right thing is often not easy. Violating a moral norm may be very attractive, especially when it involves high and immediate rewards. Internalised moral principles do not need to change this, and Gauthier therefore does not suppose that constrained maximizers face no self-benefiting opportunities. He does suppose, however, that their commitment to morality motivates them to do the right thing and refrain from violating. That commitments may enable persons to control themselves in the face of temptations is a commonly defended view both in philosophy and the social sciences (Frank, 1988; Holton, 2009).

As I mentioned in the previous section, Gauthier assumes constrained maximizers will be "affectively engaged by compliance" (p. 266). Moral sentiments such as anticipatory guilt make it easier to refrain from violating. Constrained maximizers, Gauthier suggests, have therefore reason to develop and support the development of such sentiments.[6] The same can be said about sympathy; persons who experience sympathy for others will be reluctant to exploit them. Given this function of emotions, Frank (1988) calls them *commitment devices*.

Commitments and associated emotions help the constrained maximizer to resist temptations provided by self-benefiting opportunities. The reserved maximizer, on the other hand, may expect to have more difficulty restraining himself. The reserved maximizer is not morally committed: morality is for him a set of prudential rules that he can set aside when it is in his interest to do so. He will thus not have moral sentiments such as anticipatory guilt. And, as I explained before, given that sympathetic bonding makes it difficult to take advantage of others, reserved maximizers should also have less sympathy for others. Put differently, reserved maximizers do not have the sort of commitment devices that constrained maximizers have. Of course, the reserved maximizer does have his resolution to violate only when the gains of violation are sufficiently high and the probability of detection is sufficiently low. Lacking a similar sentimental backing, such a resolution is, however, unlikely to have the same force as the constrained maximizer's commitment.

[6] In the same paragraph, Gauthier claims that persons have reason to ensure that everyone is sufficiently translucent. Given the role of emotional expressions in translucency, these points turn out not to be independent.

Reserved maximizers will be tempted to take self-benefiting opportunities. Indeed, they may experience strong visceral responses to the prospect of individual advantage such situations involve. Due to what Frank (1988) calls the psychological rewards mechanism, such responses should be particularly strong when the rewards are high and detection unlikely. Strong visceral responses cause impulses that override deliberation, and as such lead one to choose contrary to what he has planned to do (Loewenstein, 2000). Reserved maximizers will at least sometimes have difficulty controlling these impulses, particularly when the gains of violating are so high and the expected probability of detection is so low that the self-benefiting opportunity *almost* counts as a golden opportunity. What is more, we should expect they will sometimes *fail* to control themselves in the face of such impulses.[7] Put differently, they will sometimes defect when they do not have a golden opportunity.

That persons without internalised moral norms and associated emotions tend to have trouble controlling themselves in the face of self-benefiting opportunities is supported by what we know of psychopaths. Psychopaths are characterised by a lack of affective empathy and sympathy for others, have no conscience or guilt, and a disregard for morality. It is not surprising that, given such psychological properties, psychopaths have a tendency to violate norms when it is in their interest to do so. However, they also often violate norms when doing so is not in their interest, which is one of the reasons they are overrepresented in prison (see Hare, 1993). This appears to be the result of poor impulse control: impulsivity is one of the key characteristics of psychopathy (Hart & Dempster, 1997). It is also one of the criteria of the closely related antisocial personality disorder.[8] The above considerations suggest a connection between the amoral dimension and this poor impulse control: that by lacking moral commitment and associated emotions, psychopaths are more easily tempted into violating.

The main point of this subsection is that persons who intend to violate only when they have a golden opportunity will, due to limited self-control,

---

[7] This point applies of course just as well to straightforward maximizers: one should expect that, as a straightforward maximizer, one will sometimes defect when it is not in one's interest to do so. I did not explicitly incorporate this point in the previous chapter because, following Gauthier, I focussed on interactions in which it was in one's interest to violate. Incorporating this point would only reduce the expected advantage of straightforward maximization further, however.

[8] While it is now commonly believed by experts that antisocial personality disorder and psychopathy are not the same, the DSM-IV does not differentiate between the two conditions.

sometimes be moved to violate when they do not have, and may not even believe they have, a golden opportunity. This enforces the conclusion of the previous subsection: reserved maximizers are liable to also violate when the probability of detection is not in fact low.

It should be noted that, in contrast to the argument presented in the previous subsection, the argument presented here may not apply to semi-constrained maximization. Semi-constrained maximizers have internalised moral norms, and may as such also be expected to have moral sentiments that make it easier to comply. Again, this also has a downside, as such sentiments may also prevent them from taking golden opportunities. Nevertheless, semi-constrained maximizers may be expected to be less prone to violating when the probability of detection is not low than reserved maximizers.

4 The fragility of trust

The argument for reserved maximization has two premises. The first premise is that reserved maximizers do better than constrained maximizers with regard to golden opportunities. The second premise is that this gain is not offset by losses in cooperative opportunities. With regard to this second premise, the previous two sections revealed that we may expect that being a reserved maximizer affects one's cooperative opportunities negatively. Not only is the reserved maximizer less likely to be trusted than the constrained maximizer, he risks losing the trust of others by violating when the probability of detection is not in fact very low. This conflicts with the rationale behind reserved maximization. Unlike the straightforward maximizer, the reserved maximizer recognises the costs of not looking trustworthy. He seeks to violate only when doing so does not affect the trust placed in him.

In order to determine the extent of this problem for the argument for reserved maximization, this section discusses the consequences of being detected having taken a golden opportunity; in particular, how much do detected violations affect a person's perceived trustworthiness? The considerations presented in this section apply also to semi-constrained maximization.

4.1 Reputation is easily cracked…

Founding Father Benjamin Franklin supposedly said that "glass, china, and reputation are easily cracked, and never mended well." He thereby expressed what the previous chapter referred to as the trust-asymmetry principle (§3.1).

This principle states that judgments of trustworthiness are more sensitive to information suggestive of untrustworthiness than information suggestive of trustworthiness. In the previous chapter I already mentioned some findings in support of this idea. To get a better idea of how influential information of untrustworthiness may be, I shall now describe some of these findings more extensively.

In the last two decades, buying products over the Internet has become exceedingly popular. This despite the fact that, because of the separation of payment and delivery, customers are never certain that a seller will actually deliver on the agreed terms. The reason that e-commerce nevertheless works is because of the existence of what is usually called reputation systems: systems that collect, distribute, and aggregate feedback about participant's past behaviour (cf. Resnick, Kuwabara, Zeckhauser, & Friedman, 2000). Reputation systems allow buyers to choose sellers who have been observed to act trustworthily in the past, and avoid those who have been observed to act untrustworthily in the past. Consequently, having a good reputation works in a seller's favour, while having a bad reputation hurts his interest.

As the trust-asymmetry principle predicts, these effects are not equivalent: the impact of a negative reputation tends to be larger than the impact of having a positive reputation. For example, when studying the effect of sellers' reputation on the bid price of auction items at eBay, Stephen Standifird (2001) found that positive reputational ratings have a mild influence while negative reputational ratings were highly influential. Standifird looked at 81 transactions of a certain product, the Palm Pilot V, which at major brick and mortar retailers cost about $350. Sellers varied both in how much positive feedback and how much negative feedback they had received, but each of them had received more positive feedback than negative feedback. Interestingly, Standifird found that the effect of a single positive feedback comment had no significant effect on the final bid price, but that a single negative feedback comment from the past reduced the expected bid price by more than 1%. While sellers who had received ten or more positive comments could expect an increase in closing price of about 3.4%, sellers with three or more negative comments should expect a reduction of about 3.6%. Daniel Houser and John Wooders (2006) report a similar trend, finding that increases in positive comments on a seller increases a product's final price much less than increases in negative comments decrease it.

Another example comes from studies on the reputational effect of criminal convictions. There is reason to think that being convicted of a single crime

tends to have, through reputational effects, a substantial effect on one's future interest. John Lott (1992) concludes on the basis of a study of 369 drug convictions that the most significant portion of the monetary penalty imposed upon people who were convicted takes the form of reduced legitimate earnings after they return to the labour force. The main reason for this is, on Lott's view, that being convicted affects a person's reputation and through this the willingness of employers to cooperate with him. Interestingly, the overall penalty increases dramatically with the level of presentence income: while a person with a presentence income of about $20k should expect a reduction of about 10 per cent, a person with a presentence income of about $35k should expect to receive only about $20k after conviction. As a possible explanation for this, Lott suggests that higher-income convicts face a greater change in the types of jobs for which they are eligible. This fits well with the idea, presented in the previous chapter (§4.2), that cooperative arrangements with higher benefits require a higher degree of trust. This study also provides a dramatic example of how a single misstep may generate extensive distrust.

Studies such as these support the view that persons take information pertaining to untrustworthiness seriously in their interactional decisions, and more seriously than information suggestive of trustworthiness. Moreover, they underline the observation that information pertaining to untrustworthiness is unlikely to be confined to just those who observed the violation. In part due to the existence of reputation systems, such information is likely to spread. As I mentioned in the previous chapter (§3.1), studies suggest we are more attentive to information indicative of untrustworthiness, more likely to trust it, and more likely to share it with others, than information pertaining to trustworthiness (Burt & Knez, 1996; White et al., 2003).

How do people who already trust a person respond to information pertaining to untrustworthiness? This is an important question, given that the reserved maximizer is supposed to be able to retain trust relations. On the one hand, it seems likely that such information may have a particularly large impact on trusting others. As Bicchieri (2002) writes, "betrayal by an acquaintance is much more devastating than betrayal by a stranger" (p. 204). But there is also evidence that suggests prior trust may have an attenuating effect. We have a tendency to interpret new evidence in such a way that it fits our prior attitudes, which is usually called confirmation bias (e.g. Nickerson, 1998). Cvetkovich and colleagues (2002) argue this phenomenon applies to trust as well: when persons who trust a given agent are provided with

information suggestive of the latter's untrustworthiness, they discount this evidence to stick closely to their prior beliefs. In support of this, they found that the extent to which trust in the management of nuclear power plants decreased after receiving negative news about them, such as that employees were found drunk on the job, depended on how much trust they had expressed before hearing the news. Similar findings have been obtained with respect to genetically modified food (Poortinga & Pidgeon, 2004).

However, this attenuation effect should not be overstated. There is reason to think that confirmation bias may only occur provided there is space for interpretation, and not in the face of blatant violations. In line with this, Poortinga and colleagues (2004) report that some information pertaining to untrustworthiness decreased the trust of all participants, irrespective of their initial trust level. Moreover, the above studies found that the trust-asymmetry principle applied even to persons with a high level of trust in a given organisation: despite their positive attitude, information pertaining to untrustworthiness had a larger effect on their level of trust than information pertaining to trustworthiness. I conclude that although trusting persons may be willing to interpret a violation charitably if possible, in general we should expect detected violations to also negatively affect the trust of such persons.

4.2 …and never well mended

> [O]nce trust is lost, it may take a long time to rebuild it to its former state. In some instances, lost trust may never be regained. Abraham Lincoln understood this quality. In a letter to Alexander McClure, he observed: "If you once forfeit the confidence of your fellow citizens, you can never regain their respect and esteem." (Slovic, 1999, p. 697)

Reserved maximizers who have been detected violating do not have to sit still and live with the consequences. They may attempt to control the damage caused, or try to recover lost trust. This section considers how effective this should be expected to be. On the one hand, there is reason to think it may often work. For one, we know that after disappointing others it often helps to apologise. On the other hand, there is the trust-asymmetry principle. It is a corollary of this principle that lost trust is not easily regained. In addition, with Franklin and Lincoln, it is generally thought that repairing trust is sometimes not possible. I will investigate what empirical findings show about the matter. I start with discussing how effective it is to simply deny that one has taken a golden opportunity. I then discuss the effectiveness of several strategies to recover lost trust.

After having been detected violating, one may attempt to convince observers that it in fact does not count as a violation of trust. One may deny intent, or give an explanation that places the action in a less negative light. Several experimental studies suggest such explanations may be effective. In one study, participants were found to have less negative feelings towards a person who damaged their interest by failing to disclose relevant information if they were told he did so unintentionally or for altruistic reasons than if they were told he did it for himself (Shapiro, 1991). In another study, video-taped job applicants who were accused of filing a tax return form intentionally incorrectly, were perceived as having more integrity if they denied intentionality than if they acknowledged responsibility (Kim, Ferrin, Cooper, & Dirks, 2004).

In line with these findings, another study reports that victims of a harmful action were more cooperative if the perpetrator denied intent than if he acknowledged it. Gibson and colleagues let participants play a repeated Prisoner's Dilemma against another player who was, unbeknownst to the participants, controlled by the researchers (Bottom, Gibson, Daniels, & Murnighan, 2002; Gibson et al., 1999). After either 5 or 15 rounds of cooperative choices, this other player would suddenly defect. The participants would then receive a message in which the other player either acknowledged that he had defected in order to do better for himself or denied this by stating that the experimenter had made a mistake. Denials were overall no more effective than acknowledgements in convincing participants to cooperate again. However, if defection occurred after 15 rounds rather than 5 rounds, participants were more inclined to cooperate after a denial than after an acknowledgement of intentional defection. It is plausible that participants found the explanation more convincing in the light of the larger number of 'trustworthy' choices.

These studies suggest that denial of intent may, sometimes, be an effective way to control the damage of a detected violation. At the same time, there is good reason to think it will also often fail to convince. First, given that a person has in fact violated, his alternative explanation will be false. Even if this remains undetected, cooked-up explanations are in general less detailed and sound less sincere than true explanations (Vrij, 2005; Vrij, Mann, Kristen, & Fisher, 2007). They are also more likely to be thought of as inadequate (Shapiro, Buttner, & Barry, 1994). Second, several studies have found the effectiveness of an explanation to depend on the severity of the violation (Shapiro, 1991; Shapiro et al., 1994). Whereas minor violations may be easily

explained away, serious violations, which golden opportunities typically involve, require very good explanations. Third, others will often be too well informed for a denial of intent to be convincing. While participants in the above experiments had very little information to base their judgment on, in a real world setting one's interaction partners may be expected to often have information suggestive of intent. Indeed, although one of the studies just mentioned found that denying intent may preserve integrity towards uninformed observers, it also found that observers who have convincing evidence that a person violated intentionally judge him to have little integrity if he nevertheless denies intent (Kim et al., 2004).

When denial of intent does not convince, trust will be damaged. In that case, a violator may attempt to repair trust. In particular, he may attempt to convince the others that despite the violation, he is not an untrustworthy person. Trust recovery may be attempted verbally, by apologising for the transgression or by promising not to do it again. It may also be attempted nonverbally, through performing certain actions geared to recovering trust. I first discuss the expected success of the verbal approach, after which I turn to the expected success of the nonverbal approach.

To apologise for a violation is to make a regretful acknowledgement of the behaviour, often including self-castigation. By apologising one distances oneself from an action without denying responsibility for it. Apologies may thus convince others that the behaviour does not fit one's actual disposition. One may also promise not to engage in the kind of behaviour again, thereby expressing a change in intentions. While apologies for behaviour and promises not to engage in it again are distinct, they of course often go hand in hand.

I start with some positive findings. There is evidence that both apologies and promises may be effective in restoring trust. A recent meta-study shows that apologies can lead to forgiveness, a phenomenon related to trust (Fehr, Gelfand, & Nag, 2010). With regard to the effect of promises, a study by Schweitzer, Hershey and Bradlow (2006) is of particular interest for our purposes. Participants in their study played a repeated trust game with an unknown player, supposedly located in a different room. In each round, participants were given $6 and three options: pass the money to the other player, share the money in equal portions, or take it all for themselves. Passing the money on would lead to it being multiplied by 3, after which it was up to the other player to return a share of this $18 or keep everything for himself. Unbeknownst to the participants, the other player was controlled by the researchers and behaved always in the following way: he would choose

noncooperatively in the first two rounds, keeping whatever money was passed on, while in the subsequent five rounds he would choose cooperatively, returning $9 to the first player (there were 7 rounds, as participants would learn just before the 7th round). Participants were informed of the other player's choice after each round. When participants after the first two rounds were asked to rate their trust in the other player, who had just defected twice, they reported distrusting him. This was reflected in their choices: whereas in the first round over 80% of participants passed money on, only 20% of participants chose to again pass money on after these two defections. If, however, after the second round participants received a message from the other player in which he promised not to defect again and instead cooperate, over 70% of the participants cooperated in this third round.

This may look like good news for the reserved maximizer. Indeed, Schweitzer and colleagues (2006) take their findings to challenge the idea that lost trust is almost irreparable. But another finding of theirs adds a crucial qualification to this conclusion. In half of the conditions, before the first round started participants were promised by the other player that he would share the $18 if they would pass the money on. After the player violated this promise by defecting twice, subsequent promises to cooperate had no effect on the trust of participants; in the third round only 20% chose to cooperate, which did not differ significantly from when no promise was made.

Schweitzer and colleagues (2006) interpret this finding as showing that "deception may harm the trustee's credibility, and as a result subsequent promises may be viewed skeptically and be discounted" (p. 16). However, the use of the deception may not be the core issue here. Remember Bicchieri's (2006) important observation that experiments with games do not need to activate a norm to cooperate. Without specific manipulations by the researchers, there is no need for participants to believe another player *should* cooperate. However, when a player promises to cooperate such a norm *is* activated. The player's defections thereby acquire a different meaning: they reveal he is willing to break a promise, to violate a cooperative norm, and to exploit the trust of others. That is why his subsequent promises are distrusted. In that case, the finding that trust is *not* restored after the breaking of a promise is much more relevant for our purposes than the other finding. It represents the situation of a reserved maximizer who has been detected violating a cooperative norm.

Neither should a reserved maximizer expect very much from apologising. While apologising generally speaking affects forgiveness positively, the earlier

mentioned meta-study reveals two conditions that reduce the probability that a transgression will be forgiven (Fehr et al., 2010). The first condition is intent: when a transgression is intended rather than unintended it is less likely to be forgiven after apologies. A recent study has found that apologising for intentional transgressions may even *reduce* the chance of forgiveness (Struthers, Eaton, Santelli, Uchiyama, & Shirvani, 2008). The second condition is, again, how harmful the transgression is to its victims. When both conditions are satisfied, forgiveness is unlikely; as the authors conclude, "[w]hen victims perceive the offences they suffer as severe, intentional, and caused by their offenders, they are unlikely to forgive" (Fehr et al., 2010). Given that a reserved maximizer's violations are both intentional and harmful, we have reason to expect his apologies will often not be effective.

Neither promises nor apologies appear to be very effective means to repair lost trust. What about actions? Making amends appears an effective way to add weight to one's apologies, the previously mentioned study of Gibson and colleagues (2002; 1999) suggests. When the player with whom participants were playing a repeated Prisoner's Dilemma would, after having defected for several rounds, let them take the highest payoff for several rounds, participants tended to be significantly more cooperative in the final rounds of the game. In addition, they found that about 40% of participants to whom amends were made cooperated in the last round, in contrast to 20% of those who received only an apology. Similarly, Schweitzer and colleagues (2006) find that when their automatized player started cooperating after having defected twice, the cooperative choices of participants' increased substantially: whereas in the second round only 20% of the participants cooperated, 70% did so in the fifth round after having observed 3 cooperative choices of the other player.

But again, there is reason to think this will be less effective in the case of the reserved maximizer's violations. Schweitzer and colleagues find that when the other player started the interaction by making a promise that he subsequently broke, 40% rather than 70% of participants cooperate in the fifth round. In addition, when asked afterwards how much they trust the other player, these participants indicated a trust of 2.2 on a scale of 7.[9] Even through actions, then, distrust sparked by evident violations is hard to repair.

What may be more, recovering trust through amends or cooperative behaviour can only occur when given the chance. This condition is not always

[9] Participants who had not been cheated indicated a trust of 3.5. Of these, those who received both an apology and a promise indicated a trust of 4.8.

satisfied. As Schweitzer and colleagues (2006) observe, "in some settings an untrustworthy episode may lead to relationship rupture, and subsequent trustworthy behavior will be more difficult to observe" (p. 16). Indeed, it is quite likely that interaction partners whose trust has been damaged will avoid subsequent interactions. As Slovic (1999) writes, "distrust tends to inhibit the kinds of personal contacts and experiences that are necessary to overcome distrust" (p. 698).

In conclusion, a reserved maximizer's chances to, after having been detected taking a golden opportunity, control the damage or recover lost trust are not great. There are two other restrictions that should be taken into account. First, the above methods may not be effective when there is not already a solid basis of trust. One of the above studies found that denial of intent is not effective when others do not already trust a person (Gibson et al., 1999). Similarly, apologies and promises may not be expected to count for much if one a person has not already provided others with reason to trust him. Second, the above methods can only be used a limited number of times with respect to the same interaction partners. Even if a reserved maximizer is able to convince others to accept alternative explanation of a given transgression or to trust him again, their attitudes towards him are affected anyway. They may be on their guard, with an increased attentiveness for new information about untrustworthy behaviour. Consequently, detected violations reduce the probability of getting away with future violations.

5 Implications for reserved maximization

The conclusion of the previous two sections is that reserved maximizers will at times incur serious and possibly irreparable damage to their perceived trustworthiness. What are the implications of this conclusion for the rationality of reserved maximization? The rationale for reserved maximization was that it combines the best elements of straightforward and constrained maximization. Like straightforward maximizers, reserved maximizers can occasionally exploit cooperative arrangements to their own advantage. But as they usually choose cooperatively with respect to self-benefiting opportunities, they were supposed to also generate the trust required to have similar cooperative opportunities as constrained maximizers. The implication of the above is that this strategy is unlikely to be successful. By violating, and in particular by violating when the probability of detection is not in fact low, reserved maximizers will increase their translucency and generate the

impression they are untrustworthy. They will occasionally destroy trust relations they have slowly developed, and reduce their attractiveness as new interaction partners. Even if they do not fully lose the trust of others, they will look less trustworthy than constrained maximizers, and as such be less likely to gain access to those fruitful cooperative arrangements that require a particularly high degree of trust. In short, reserved maximizers do less well with respect to cooperative opportunities than constrained maximizers. It is far from evident whether this is compensated by the benefits reaped from golden opportunities. As such, it is far from evident that the second premise of the argument for reserved maximization is correct.

There may be a way to solve this problem. Reserved maximizers can become *more* reserved. The reserved maximizer needs to somehow take into account his limitations. He needs to take into account that an apparent golden opportunity may be *merely* apparent and that exploiting it could seriously hurt his interest. Can this be done? One way to do so would be by violating *only* when the benefits involved in doing so are so high that they would compensate the loss in cooperative opportunities that would result *if* the violation were detected. Such a *very reserved maximizer* would not violate whenever he believes he has a golden opportunity, but only when the expected benefits of defecting outweigh its potential costs. This change can easily be incorporated in the original description of reserved maximization. The reserved maximizer was described as defecting when the probability of detection is very low and the difference between the initial payoffs of exploitation and cooperation is above a certain threshold value. The thrust of this proposal is to increase this threshold value so that it exceeds the potential costs of being detected violating.

The very reserved maximizer should be substantially less likely to hurt his future interests by violating than the original reserved maximizer. But two other problems remain. First, this adaptation does not take care of limited self-control. Very reserved maximizers will also be tempted to defect by self-benefiting opportunities that do not satisfy their higher threshold. While the stricter resolution may be of some help, they would still lack the compliance supporting sentiments that constrained maximizers may have. Second, the increase in threshold should not be expected to change the fact that already before being detected violating reserved maximizers looks less trustworthy than constrained maximizers. The dispositions are still associated with different behaviours: a very reserved maximizer will also be calculating, lack

genuine emotional expressions, and rely on manipulation and deception. Put differently, a very reserved maximizer would still be translucent.

This poses a problem for very reserved maximization. Again, it is unclear exactly how translucent very reserved maximizers are due to these differences. It is not unlikely that this degree of translucency is so low that they would only look a little less trustworthy than constrained maximizers do. Nevertheless, the associated loss in cooperative opportunities may be sufficient for constrained maximization to be more advantageous. Due to their high threshold, very reserved maximizers will only rarely exploit cooperative arrangements to their advantage. Even if these rare golden opportunities are very advantageous, they may not be sufficient to make up for the reserved maximizer's loss in cooperative opportunities. I therefore conclude that it is not evident that reserved maximization, including its very reserved subtype, is more advantageous than constrained maximization.

The case for semi-constrained maximization may be stronger. As the conclusions of the previous two sections also apply to this disposition, a semi-constrained maximizer should also become very reserved and will in effect reap substantially smaller benefits from taking golden opportunities. However, there is an argument to be made that, due to having similar sentiments as constrained maximizers, semi-constrained maximizers will look more trustworthy than reserved maximizers. Not only will their emotional expressions be more genuine, they will also be less tempted to violate when they do not have a golden opportunity. If that is correct, semi-constrained maximization is a more advantageous disposition than very reserved maximization. It is as such also more likely to trump constrained maximization. On the other hand, there remains the aforementioned counterargument: sentiments such as (anticipatory) guilt may make it both more difficult to take golden opportunities and to hide having taken them, and may thus also increase the semi-constrained maximizer's translucency.[10] As I do not know how to decide this matter, my conclusion is that, provided semi-constrained maximizers are able to control their moral sentiments with respect to golden opportunities, or at least their expression, semi-constrained maximization has a stronger case than very reserved maximization.

---

[10] It may be argued that as a semi-constrained maximizer believes he should take golden opportunities, he will not be burdened by guilt. It is unclear, however, whether the *pro tanto* reason against doing so may not nevertheless evoke such an emotion. Do we not often feel bad about decisions that we believe to be all things considered the best decision?

Before moving on, it is worth noting that defence of constrained maximization diverges from Gauthier's approach. As I mentioned before, Gauthier does not just hold that it is rational for us to be moral, but that morality reduces to rationality. He argues as such that fully rational persons would be constrained maximizers. The argument for constrained maximization, Gauthier emphasises, is therefore not supposed to rely on the imperfections of human agents: "The rationale for disposing oneself to constraint does not appeal to any weakness or imperfection in the reasoning of the actor; indeed, the rationale is most evident for perfect reasoners who cannot be deceived" (p. 186). My argument in this chapter, in contrast, did appeal to certain imperfections. In particular, I have argued that we must take into account that human agents are imperfectly informed, that they reason imperfectly about risks, and that they have limited self-control. Given that these imperfections are facts about our condition, they must be taken into account when we consider the expected advantage of reserved maximization for *us*. It is important to emphasise, however, that my argument does *not* apply to agents without these imperfections.

6 Reserved maximization and individual differences: the case of the successful psychopath

As translucency depends on properties that vary between individuals, some persons have a lower degree of translucency than others. Dispositions such as straightforward maximization and reserved maximization will be more advantageous to such persons than for a more translucent person. In the previous chapter I have argued that even for the less translucent it is not advantageous to adopt straightforward maximization in favour of constrained maximization. This seems not to be the case for reserved maximization. It is not implausible, I shall argue below, that there is a type of person for whom reserved maximization is more advantageous than constrained maximization.

My argument that reserved maximization is not evidently more advantageous than constrained maximization depends on persons having certain psychological limitations. While it is plausible that these psychological limitations apply to everyone to a certain extent, it is also plausible that there are individual differences. Some persons are better at lying, manipulating, faking their feelings and controlling their impulses than others. As I mentioned before (7§2.4), a meta-analysis on lie-detection has found that some people are substantially better at lying than others. It is not hard to see that

for persons who have extraordinary capacities for deceiving and manipulating others, as well as the ability to control their impulses, reserved maximization may be more advantageous than constrained maximization.

The crucial question is whether there is reason to expect that such 'deceptive people', as Sayre-McCord calls them, may in fact exist. Popular culture suggests they do. There are many fictional characters with great deceptive skills. Frank Abagnale Jr., who is portrayed in *Catch Me If You Can*, is able to con (almost) everyone into believing he is a doctor, a pilot, or a lawyer—even though he has not been schooled in any of these professions. On a more serious note, Verbal Kint from *The Usual Suspects* and Tom Ripley from *The Talented Mister Ripley* are con artists who are so skilled at deception, manipulation and impersonation that they are literally able to get away with murder. Another example of an exceptionally skilled deceiver is Dexter Morgan, the main character of the popular TV-show *Dexter*. Dexter is able to have a relatively normal life as a friendly and trustworthy-looking blood spatter analyst while being a prolific serial killer in his off hours.[11] These fictional characters appear to have the capacities to be successful reserved maximizers.

That there are fictional characters who are sufficiently deceptive that they can make reserved maximization successful does not of course imply that there are actual persons for whom this is the case.[12] However, the psychological profile that fits most if not all of these characters is one that we can find also outside of fiction. Kint, Ripley, Morgan and possibly Abagnale have many of the characteristics associated with psychopathy. Psychopathy is characterised by traits such as deceitfulness, fearlessness, disregard for others, lack of remorse or guilt, lack of empathy, a superficial charm, grandiosity and manipulativeness.[13] Importantly, clinical studies of psychopaths do not only

[11] Although he continuously lies to those who trust him in order to hide his 'Dark Passenger', Dexter is to a certain extent trustworthy: generally speaking, he sincerely tries to make good on his commitments to others. When he does violate norms by killing others, it is in relation to persons whom themselves have done so as well.

[12] Frank Abagnale Jr. does in fact exist. His career as a con man has not been unequivocally successful, however, as he has been in prison several times.

[13] Psychologists tend to describe psychopathy partly in terms of how they behave. Hare, the renowned expert on psychopathy, describes them for example as "social predators who charm, manipulate, and ruthlessly plow their way through life… Completely lacking in conscience and feeling for others, they selfishly take what they want and do as they please, violating social norms and expectations without the slightest sense of guilt or regret" (1993 p. xi). Psychopaths are in this case already defined in terms of how they choose with regard to moral situations. Conceptually, however, I think we can—and for present purposes should—distinguish between a psychological profile and a disposition to choose. When I speak of a psychopath, I mean to

show that psychopaths deceive and manipulate a lot, but suggest also that they tend to have skill in doing so (e.g. Hare, 1993). While there is little experimental work regarding this, a recent study finds that psychopathic traits are associated with a better ability to control insincere emotional expressions and decrease leakage of one's genuine feelings (Porter, Brinke, Baker, & Wallace, 2011). As the researchers suggest, this may be precisely because of the psychopath's emotional shallowness: lacking real feelings for others, there may be less genuine emotion to interfere with what one wants to express.

In line with their reputation in popular culture, several researchers have pointed out that psychopaths appear to have the sort of characteristics that may enable one to be a successful cheater in a population of cooperators; indeed, some theorist have attempted to explain psychopathy as an evolutionary adaptive strategy (e.g. Harpending & Sobus, 1987). These characteristics are also clearly useful for reserved maximizers. A person with an enhanced ability for deception and manipulation is more likely to get others to trust him in the first place, and more likely to contain the damage of detected violations.

While they may have the requisite skills of deception and manipulation, there is reason to think that most psychopaths do not have sufficient self-control to be successful reserved maximizers. As I mentioned before, besides the abnormal affective and interpersonal traits described above, psychopaths are typically more impulsive than non-psychopaths. They are often also more aggressive and irritable, with the result that they are more likely to get into fights and fail to conform to social norms even when it is clearly in their interest to do so. They tend to be unable to maintain enduring relationships and often end up in prison, psychiatric institutions, or on the street (Coid, Yang, Ullrich, Roberts, & Hare, 2009; Comer, 2003).[14] They also tend to form unrealistic life-plans, and to fail to stick to the plans they make (Hare, 1993). This suggests that psychopaths would have more difficulty in violating only when actually facing a golden opportunity than non-psychopaths.

But what if there would be persons who have those psychopathic traits that support reserved maximization, but not those traits that increase the

---

refer to a person with the psychological profile of a psychopath, without implying that this person pursues an amoral strategy.

[14] Psychopaths who get caught may also be able to use their skills effectively to restore trust. Incarcerated psychopaths are more able to feign remorse through which they can get lower sentences, and it has been reported that psychopaths are 2-and-a-half times as likely to be released when they apply for parole than non-psychopaths (Porter, Ten Brinke, & Wilson, 2009).

probability of detection? It has become common to dissociate between the affective-interpersonal dimension of psychopathy and the impulsive-antisocial dimension (Porter et al., 2011). In the past decade, more and more theorists of psychopathy have come to believe there is a subtype of psychopathy that includes the affective-interpersonal dimension of psychopathy but not the lack of self-control. Persons with such a psychological profile may use their skill for manipulation and deception to get ahead in life, while having sufficient control over their impulses to avoid ending up in prison. In their book *Snakes and Suits*, Paul Babiak and the aforementioned Robert Hare (2006) argue that these so-called 'successful' psychopaths may be occupied as politicians, lawyers, CEOs, or professors. They describe them as looking like normal, trustworthy people, who usually also act as such, but who are in fact fearless, cunning, and remorseless beings that do not let morality stand in the way of personal gain. Indeed, several famous or infamous persons have, albeit without the requisite tests, been 'diagnosed' as successful psychopaths, including Sir Richard Burton, Churchill, Charles Yeager, President Lyndon Johnson, the Enron executives Kenneth Lay and Andrew Fastow, and Bernard Madoff (e.g. Hall & Benning, 2006).

Although it is generally thought that successful psychopaths exist, little is known about them. Participants in studies on psychopathy are usually incarcerated, and thus per definition not very successful. Furthermore, there are serious difficulties in obtaining suitable subjects: advertisements for volunteers for a study of successful people who are psychopathic do not get many responses. Stephanie Mullins-Sweat and colleagues (2010) thus tried something different. They reasoned that even though successful psychopaths are unwilling to participate in studies, individuals closely familiar with them may be able to provide information about how they function. They questioned individuals in professions likely to come into contact with psychopathic individuals whether they know anyone whom they would describe as a successful psychopath and, if so, to describe this person in terms of traits associated with psychopathy. They found that many of the respondents believed they were familiar with a successful psychopath. Interestingly, while these individuals were characterised as high in typical psychopathic traits such as callousness, dishonesty, and as being exploitive and remorseless, they were not thought of as being impulsive or irresponsible. Instead, they received high scores on traits associated with conscientiousness, such as discipline, competence, achievement-striving, and deliberation.

Babiak, Hare and collaborators (2010) have investigated successful psychopathy more directly. They were given the unique opportunity to examine psychopathy and its correlates in a sample of 203 corporate professionals selected by their companies to participate in management development programs. Scoring each individual on a psychopathy checklist, Babiak and colleagues (2010) found that about 3 per cent of them scored above the common research threshold for psychopaths.[15] Most of the participants with high psychopathy scores held high-ranking executive positions: they were vice-presidents, supervisors, and directors. In line with the profile of the successful psychopath, high scores on the psychopathy test were associated with high in-house ratings on a scale for charisma/presentation style, including items for creativity, communication skills, and strategic thinking. Most interestingly, the researchers found that these persons managed to hold their powerful positions even though their immediate bosses rated their performance and their management styles to be relatively low and considered them to be bad team players. Babiak et al take this to be evidence of the successful psychopath's capacity to manipulate others into trusting them in the face of contrary evidence.

These studies suggest, then, that successful psychopaths do exist. Not only do such persons have capacities that are helpful for reserved maximization, they appear to live in a way that resembles reserved maximization. Moreover, there is an argument to be made that in their case this disposition is more advantageous than constrained maximization. First, they are likely to be better at reserved maximization than most of us. They have the deceptive and manipulative skills that enable them to convince others of their trustworthiness, even in the light of contrary evidence. Second, and just as importantly, they may be *worse* at constrained maximization than most of us. Normal moral agents can and do develop social and moral emotions that make it easier for them to function as constrained maximizers. Psychopaths, on the other hand, may well be unable to have these sorts of emotion. This does not only mean that they will have more difficulty complying in the face of golden opportunities, it also means that they will not automatically display the kind of emotions that make them attractive interaction partners. To be successful as constrained maximizers, or semi-constrained maximizers for that matter, they must thus engage in some sort of deception, pretending to have

[15] As a recent study found a prevalence of 0.6% in the general population (Coid et al., 2009), this finding supports the claim that successful psychopaths are more attracted to certain environments than others.

emotions they do not have. As this will be difficult, they will look less trustworthy as constrained maximizers than most of us, and possibly just as untrustworthy had they been reserved maximizers. If so, they are better off as reserved maximizers.

Psychopaths typically do not take themselves to have reason to act morally. The above suggests that, at least if we accept Gauthier's approach towards morality, they may sometimes be right. For some psychopaths it is not rational to choose constrained maximization over reserved maximization and thus not rational to be moral.

7 Conclusions

Many theorists have criticised Gauthier's claim that constrained maximization is the most advantageous disposition to have in one's interactions with others. This chapter considered another contender for this title, reserved maximization. The argument for the rationality of reserved maximization depends on two premises. First, reserved maximization is a more advantageous disposition to have with regard to situations that involve golden opportunities. Second, this gain with respect to golden opportunities is larger than potential losses in cooperative opportunities. This chapter targeted the second premise, and concluded it is not evidently satisfied.

Drawing in part on findings regarding translucency from the previous chapter, I concluded first that reserved maximizers should be expected to look less trustworthy than constrained maximizers. They are more calculating, have shallower emotional bonds with others, and lack moral sentiments. In addition, they must engage in manipulation and deception, both in order to get golden opportunities and to avoid being detected having taken them. Reserved maximizers are therefore translucent *even* when they have not been detected violating. On the assumption that people prefer not to cooperate with persons who are untrustworthy in the way reserved maximizers are (§1), this means that being a reserved maximizer affects one's cooperative opportunities negatively.

Translucency increases substantially when violations are detected. This is likely to occur now and then, I argued, as reserved maximizers will sometimes violate when they in fact do not have a golden opportunity. Due to limited information about the probability of detection and psychological biases, persons are prone to over-detect golden opportunities. Furthermore, without moral commitment and associated emotions, reserved maximizers will

sometimes be tempted into violating even though they do not believe they have a golden opportunity.

Being detected taking a golden opportunity will seriously hurt the reserved maximizer's cooperative opportunities. Empirical studies confirm the common sense idea that trust is fragile: it is easily lost and may never be regained. To secure their interests, reserved maximizers must become very reserved and let most golden opportunities pass. This does not solve the problem that they will now and then be unable to resist the temptation to defect, however. Nor does it change the fact that reserved maximizers look less trustworthy than constrained maximizers. It is not unlikely that reserved maximizers' loss in cooperative opportunities is not compensated by the benefits associated with occasionally reaping a golden opportunity. It is therefore not evident that reserved maximization, including its very reserved subtype, is more advantageous than constrained maximization.

That is the good news for Gauthier's argument for constrained maximization. There are, however, also two important caveats. First, there appears to be a group of persons for whom it is true that they are better off as reserved than as constrained maximizers. Successful psychopaths appear to have the capacities to get away with reserved maximization. In addition, due to emotional abnormalities, constrained maximization may be less profitable to them than for others.

Second, another disposition, semi-constrained maximization, may be more advantageous than both reserved and constrained maximization. Unlike the reserved maximizer, the semi-constrained maximizer takes himself to have reason to act morally; unlike the constrained maximizer, this reason can be overruled by reasons provided by his interests. Although much of the above also applies to semi-constrained maximizers, they have moral commitments and associated sentiments like constrained maximizers and may therefore look more trustworthy than reserved maximizers. Provided that these same moral sentiments create difficulties with respect to either taking golden opportunities or hiding haven taken them, having this disposition is less costly to one's cooperative opportunities than reserved maximization. Semi-constrained maximization would in that case be more advantageous than reserved maximization, and be also more likely to surpass constrained maximization.

9

When Constrained Maximization is Rational

1 Introduction

Gauthier's answer to the 'Why be Moral?' question is that being moral is the best way to satisfy one's amoral aims and desires: that it maximizes one's expected utility. His argument for this claim relies on the uncontroversial idea that people prefer to interact with persons who can be trusted to comply with moral constraints, and the more controversial idea that people can recognise whether persons are so disposed or not—that persons are translucent. The acceptance of moral constraints may therefore be expected to have a positive effect on one's cooperative opportunities, Gauthier claims. More precisely, Gauthier claims that people are to such a degree translucent that being moral is more advantageous than not being moral. I have investigated this Translucency Assumption in the previous two chapters. In this chapter I discuss the implications for Gauthier's argument for constrained maximization. I start by laying out the reasons for regarding the Translucency Assumption as plausible. Section three discusses several ways in which we may increase the expected advantage of constrained maximization, thus giving further credence to the Translucency Assumption.

2 What the findings do and do not show

The previous two chapters have compared the disposition of constrained maximization with two amoral alternatives, straightforward maximization and reserved maximization, in the light of findings on translucency. The conclusion of Chapter 7 is that people are to such a degree translucent that constrained maximization is more advantageous than straightforward maximization. Although it presented a greater challenge, the conclusion of Chapter 8 is that it is far from evident that reserved maximization is more

advantageous than constrained maximization. The three challenges that were introduced in Chapter 6 (§5) have thereby been deflected.

The case for constrained maximization in comparison with these two other dispositions is in fact even stronger. The above chapters concentrated on the benefits and costs that having a certain disposition has for one's cooperative opportunities. These are, however, not the only benefits and costs that we may take into account when considering the expected utility of a disposition. I now want to mention two other benefits that further strengthen the case for constrained maximization against these two other dispositions.

First, there is reason to think that most persons are sociable in such a way that they value standing in relations of mutual trust for its own sake. Friendships and other trust relations are valuable to the great majority of us, and not just for the cooperative opportunities they provide. Straightforward maximizers are unlikely to get such benefits.[1] Due to their untrustworthiness they are unlikely to develop trust relations or to enter trust communities. The social relations they will have with others should thus be expected to be impoverished. With regard to reserved maximization this is less clearly the case. Reserved maximizers can develop trust relations with others and function in trust communities, even though the trust of others in them would be misplaced. However, these relations may still be impoverished in certain ways, as such relations will not, at least from the reserved maximizer's side, involve the commitment and emotional involvement typical of trust relations. Therefore, persons who care about their relations with others have another reason to choose constrained over straightforward maximization, and possibly also over reserved maximization.

Second, there is reason to think that constrained maximization has certain benefits with regard to cognitive processing over the other dispositions. Constrained maximizers have internalised moral norms. When faced with self-benefiting opportunities, the only thing constrained maximizers may need to figure out is whether others can be trusted not to exploit them. Straightforward and reserved maximizers, on the other hand, need to calculate the difference in costs and benefits between cooperating and violating on a case-to-case basis. Reserved maximizers, in particular, seem to have to do some cognitive heavy-lifting in order to determine whether their threshold is reached or not. This counts also for very reserved maximizers, and thus

[1] Note that this point was not taken into account in Chapter 7; while I claimed that trust would affect payoff matrices by affecting the risks parties are willing to take, this aspect of trusting relations was not included in the payoffs.

provides an additional reason to think persons are better off by internalising moral principles.

Besides these two amoral dispositions, the previous chapter considered a third alternative to constrained maximization more briefly. Semi-constrained maximizers have internalised moral norms, and take themselves to have *pro tanto* reason to comply with these norms; but when the benefits of violating are very high they take themselves to have stronger reason to violate. Like reserved maximizers, semi-constrained maximizers may expect to occasionally profit from golden opportunities. There is, however, an argument to be made that semi-constrained maximizers may expect to get better cooperative opportunities than reserved maximizers: given that they have internalised moral norms, and may expect to have associated moral sentiments, they look more like constrained maximizers. If that is correct, semi-constrained maximization has a higher expected advantage than reserved maximization. It is therefore also more likely to be more advantageous than constrained maximization. This argument in favour of semi-constrained maximization does, however, need an assumption that is not evidently satisfied: that the semi-constrained maximizer's moral sentiments, including guilt, do not intervene with his abilities to take golden opportunities and to hide having taken them. I therefore conclude it is not evident that semi-constrained maximization is more advantageous than constrained maximization.

That is not to conclude, however, that the Translucency Assumption is fully in line with empirical findings. Not everyone has reason to adopt constrained maximization in favour of reserved maximization. As was discussed before (6§5.2, 7§3.3 and 7§4.2), Sayre-McCord (1991) has challenged Gauthier's argument for constrained maximization by pointing out that there may be persons who are sufficiently skilled at deception that they are better off as straightforward maximizers. Although I have argued that there is no reason to think there actually are persons for whom this is true, there is a minority for whom reserved maximization may be expected to be more advantageous than constrained maximization. Persons with the psychological profile of a successful psychopath appear to have the skill set required to make reserved maximization work. In addition, as they lack moral sentiments and a conscience, they should expect to do worse as constrained maximizers than other persons. It is not clear whether the two additional arguments in favour of constrained maximization mentioned affect this conclusion. Persons with the psychological profile of the psychopath appear not to value social relations

in the way that most of us do. Processing costs may be relevant, but may not carry sufficient weight.

Given this caveat, we may not conclude that everyone has reason to adopt constrained maximization. But for the great majority, neither reserved maximization nor semi-constrained maximization is evidently more advantageous than constrained maximization. The assumption that persons are sufficiently translucent such that, for the great majority, constrained maximization is more advantageous than reserved or semi-constrained maximization is thus compatible with empirical findings.

That this weaker variant of the Translucency Assumption is empirically plausible is an important result for Gauthier's defence of the rationality of morality. Gauthier uses the argument for constrained maximization to show that when a set of norms is the object of a rational and general agreement, it is also rational to comply with these. The weak Translucency Assumption allows him to claim that would this agreement be in place, it is rational for most of us to be disposed to comply with it when others are similarly disposed. Furthermore, he may use it to defend his claim that, in so far as actual norms approximate these ideal norms, it is rational for most of us to be disposed to comply with these.

This last point reveals that the result is also relevant for moral theory more generally. Constrained maximization may be advantageous even when the generally accepted norms are different from those that would, on Gauthier's view, be the object of rational agreement. As I said before: what matters most is not the exact content of the accepted norms, but that being disposed to comply with them yields cooperative opportunities (6§3). Because the argument for constrained maximization is not tied to Gauthier's particular moral conception, other moral theorists may also use it to address questions of motivation. Self-interest is an important motivator, and the argument for constrained maximization may thus be motivationally efficacious. Of course, the conclusion that not everyone is better off as a constrained maximizer means also that the argument will not be effective for everyone.

However, my findings do fail to support another important aspect of Gauthier's project. As I described in the introductory chapter, Gauthier does not merely want to show that for most actual persons it happens to be rational to be moral; he seeks to reduce morality to rationality (§2.2). Or as he puts it, to show that "to choose rationally, one must choose morally" (1986, p. 4). That there are persons who are better off as amoral reserved maximizers is inconsistent with this reductive claim. As these persons would rationally

choose this disposition, they would also, on Gauthier's own terms, rationally choose immorally.

That Gauthier's reductive claim is unsupported becomes even clearer once we turn from the question of what disposition actual persons would rationally choose to the question of what *idealised* agents would rationally choose. Gauthier defends his reductive claim by arguing that fully rational persons would agree to and comply with the contractarian morality he defends. As I mentioned before (8§5), to let his argument avoid depending on human limitations, he explicitly assumes these persons to be perfect reasoners without our weaknesses and imperfections. My findings do not suggest, however, that such persons would be sufficiently translucent to adopt constrained maximization (cf. Den Hartogh, 1993). In particular the argument against reserved maximization depended on the assumption that agents have certain imperfections or weaknesses. It is due to incomplete information, a tendency to reason imperfectly about risks, and limited self-control that we are unable to reliably detect and act on golden opportunities. A being without these limitations, on the other hand, should not have this problem. Such a being, then, may rationally choose to be a reserved maximizer rather than a constrained maximizer.[2]

While moral commitment is instrumentally rational for beings as imperfect as ourselves, it may not be for perfect beings. Indeed, my findings suggest that it is rational for us to be moral in part because we are not good enough at being amoral.

3 How to get the most out of constrained maximization

As Gauthier points out at several places in *Morals by Agreement*, whether persons are sufficiently translucent for constrained maximization to be rational depends in part on factors that are in their control. In this final section I discuss three types of methods by which constrained maximizers can increase the expected advantage of their own disposition and decrease the expected advantage of alternative dispositions. The types of methods are mentioned by Gauthier; I shall elaborate on them where relevant on the basis of my present investigation. Some of these methods played a crucial role in my analysis, and are as such not optional: they must be adopted for constrained maximization to be more advantageous than straightforward and reserved maximization. I will

[2] Transparent ideal agents *would* choose constrained maximization over reserved maximization. But it is unclear why we should assume that perfect reasoners are transparent.

also suggest methods constrained maximizers *may* adopt to increase their expected advantage with respect to alternative dispositions further. These dispositions include very reserved and semi-constrained maximization, dispositions of which I have argued that they are not evidently less advantageous than constrained maximization. By adopting the method below, constrained maximizers can thus increase the rationality of their disposition.

3.1 Strive for optimality and against unfairness

Gauthier's argument for constrained maximization is supposed to show that when a community of rational persons agrees on cooperative arrangements that are fair and optimal (i.e. that lead to Pareto efficient outcomes), it is also rational for them to comply with these arrangements. Gauthier insists, however, that the argument for constrained maximization also applies when arrangements "fall short of the ideal" and are only "nearly fair and optimal" (p. 168). Nevertheless, constrained maximizers do have reason to support developments that bring arrangements as close as possible to the ideal. Consider first optimality. When arrangements are not optimal, the fruits of cooperation are not as high as they could be. This also affects the degree of translucency required for constrained maximization to be advantageous. As Gauthier writes, "as practices and activities fall short of optimality, the expected value of co-operation [...] decreases, and so the degree of translucency required to make co-operation rational increases" (p. 178). This was made explicit in Chapter 7 (§2.5 and §4.2): the larger the gain from cooperating with others over noncooperation, the larger the expected advantage of constrained maximization, and the lower the degree of translucency required for the Translucency Assumption to be satisfied.

Constrained maximizers may not benefit directly from increasing the fairness of arrangements in the way that they benefit from increasing optimality. Increases in fairness should, however, benefit them indirectly by reducing the likelihood of becoming exploited. Individuals who are treated unfairly under a given arrangement are less likely to comply with it. As numerous experiments have confirmed, people have an deep-seated aversion to being treated unfairly. Experiments with the Ultimatum Game show that people reject unfair treatment by others even if they would be better off by accepting it (Oosterbeek, Sloof, & Van De Kuilen, 2004). Moreover, it may not even be rational for persons who are treated unfairly to comply. As Gauthier writes, "as practices and activities fall short of fairness, the expected value of co-operation for those with less than fair shares decreases, and so the degree of

translucency to make co-operation rational for them increases" (p. 178). Like increases in optimality, increases in fairness may thus increase the expected advantages of constrained maximization for others, and thereby decrease the degree of translucency required for the Translucency Assumption to be satisfied.

Although the fairness and optimality of norms did not play an explicit role in the previous chapters, it did implicitly: the analysis assumed that each interaction partner in a cooperative exchange stands to gain substantially over not cooperating. In that case, the arrangement cannot be wholly unfair and suboptimal. But cooperative arrangements may likely be improved beyond what my argument required. Constrained maximizers may strive for such improvements both in their individual interactions, and support improvements on a societal level. This does not only benefit them directly, but also increases the loss of social exclusion that other dispositions risk.

Besides supporting the development of fair and optimal cooperative arrangements in general, constrained maximizers should also act strongly against being treated unfairly themselves. Gauthier gives two arguments for this. The first is that it is in a person's interest to be so disposed. This argument appeals to translucency: "for in so far as she is known to be broadly compliant, others will have every reason to maximize their utilities at her expense, by offering 'co-operation' on terms that offer her but little more than she could expect from non-co-operation" (p. 178). In view of the previous chapters, this assumption of translucency is plausible. Especially in an informed setting, others are not unlikely to become aware of whether they can use you or whether you will stand up for yourself. The second argument is that by doing so the constrained maximizer "ensures that those not disposed to fair co-operation do not enjoy the benefits of any co-operation, thus making their unfairness costly to themselves, and so irrational" (p. 179). By refraining from cooperating with people disposed towards unfair behaviour one can reduce the expected advantage associated with such dispositions. I shall have more to say about this in the next subsection.

3.2 Trust cautiously

By striving for optimality and against unfairness, constrained maximizers can increase the expected advantage of their disposition. They should however also adopt methods to increase the costs of not being so disposed. One such method was just mentioned: by refraining from cooperating with unfair others, one may increase the costs of being unfair. But they can and must do more.

Gauthier writes that "we should not suppose it is rational to dispose oneself to constrained maximization, if one does not also dispose oneself to exclude straightforward maximizers from the benefits realizable by co-operation" (p. 180). It is not hard to see why. Constrained maximizers hurt their interest if they do not keep amoral others out of cooperative arrangements. Failing to do so creates the risk of being exploited. Not only this, "it ensures that the arrangements will prove ineffective, so that there are no benefits to share" (p. 180). By excluding untrustworthy others from cooperative arrangements, constrained maximizers also increase their expected advantage indirectly, namely by reducing the prospect of being untrustworthy in the first place.

The expected advantage of constrained maximization as well as the expected advantage of the other dispositions thus depends crucially on how able constrained maximizers are at excluding untrustworthy others from cooperative arrangements. So does the plausibility of the Translucency Assumption: the better constrained maximizers are at excluding untrustworthy others, the less translucency is required for constrained maximization to be rational. Although this already played a crucial role in Gauthier's argument, my analysis of the previous chapters implies that constrained maximizers must take it a step further.

Gauthier defines the constrained maximizer as cooperating only when she expects her interaction partners do so as well. 'Expect' is an ambiguous term, however. It may be taken to mean that she cooperates whenever she deems the probability that others will cooperate as well to be more likely than that they do not. But the previous chapters suggest that constrained maximizers must be more cautious than this. My argument that constrained maximization is more advantageous than straightforward maximization depended on the assumption that constrained maximizers interact predominantly with others they believe very likely to cooperate (7§4), as did my argument for why both reserved and semi-constrained maximizers must be very careful with respect to golden opportunities (8§5). I assumed constrained maximizers avoid giving self-benefiting opportunities not only to persons they *distrust*, but also to persons they *do not* trust. In addition, I assumed constrained maximizers make their willingness to cooperate dependent on the stakes of interactions. While they may take small risks with persons whom have not yet proven themselves trustworthy, they will only make themselves vulnerable when interacting with others with whom they have developed trust relationships. Due to this cautiousness of constrained maximizers, persons with alternative dispositions

will gain less from cooperative opportunities merely by looking less trustworthy than constrained maximizers.

In addition, I assumed constrained maximizers form their expectations about the trustworthiness of others in accordance with the trust-asymmetry principle: that they take information pertaining to untrustworthiness more seriously than information pertaining to trustworthiness (7§3, 8§4). It is due to this principle that being detected violating is likely to have serious consequences for one's future interactions. More precisely, it is partly due to this principle that straightforward maximizers are substantially more translucent in informed interactions than in isolated interactions, that straightforward maximizers are unlikely to develop and retain trust relations, and that reserved and semi-constrained maximizers must only seldom reap golden opportunities. This principle being generally accepted is crucial for constrained maximization being more advantageous than alternative dispositions.

My analysis thus assumes that constrained maximizers are quite cautious with respect to the conditions under which they give others an opportunity to take advantage of them. But they can be more cautious than this, and thereby decrease both the risk of being exploited and the expected advantage of alternative dispositions further. Provided they have developed sufficient trust relations, they may, instead of only interacting predominantly with persons who have proven themselves trustworthy, *exclusively* interact with such persons. In addition, they may operate on an even more asymmetric trust-asymmetry principle, and take information pertaining to untrustworthiness extremely seriously, as by operating on a 'one strike and you're out' policy.

Of course, there is a limit to how cautious constrained maximizers should be. Cautiousness should not lead one to constant distrust. Such paranoia will not only come with psychological costs, but will also make it difficult for one to develop and maintain trust relationships. Cautiousness should also not lead to an undue reliance on stereotypes, even if there is a grain of truth in them. Such behaviour implies a kind of unfairness which, we saw, is not to the advantage of constrained maximizers.

3.3 Increase translucency

The third and most obvious type of method that constrained maximizers can adopt to increase the expected advantage of their own disposition is to simply increase translucency. In particular, they should enhance their ability to detect whether others are trustworthy or not. As Gauthier writes: "those who believe

rationality and morality to be at loggerheads may have failed to recognise the importance of cultivating their ability to distinguish sincere co-operators from insincere ones" (p. 181). By becoming better at detecting the dispositions of others, constrained maximizers increase the chances of achieving mutual cooperation with other constrained maximizers and reduce the chances of being exploited by straightforward, reserved, or semi-constrained maximizers. At the same time, it reduces the benefits that persons with these alternative dispositions may expect from both exploiting as well as cooperating with constrained maximizers. The previous chapters led to a few suggestions regarding what constrained maximizers may do to increase translucency.

Individuals may improve their individual ability to detect the trustworthiness of others. For one, it has been argued that we can become better at detecting lies (Vrij, 2004). Although training programmes do not always improve accuracy (Akehurst, Bull, Vrij, & Köhnken, 2004), there are some positive findings (Vrij et al., 2008; 2011a). In particular, it appears that people can become better lie detectors by increasing the cognitive load of those whom they are judging, for example by asking them difficult questions or by asking questions when they are distracted(Vrij, Granhag, Mann, & Leal, 2011b). In addition, people can learn about the many false assumptions that reduce lie detection accuracy, such as that there are certain unique facial cues associated with lying or that people who lie are typically more nervous (Vrij et al., 2011a).

The probability of detecting another person's disposition increases when observers are more familiar with the person's past behaviour or when third party judgments regarding his trustworthiness are available. My analysis assumed constrained maximizers therefore ensure their interactions usually satisfy at least one of these conditions (7§4). But agents can increase the translucency of their interaction partners further by ensuring that both conditions are met most of the time. If we take existing persons as a reference point, advances surely seem possible. Most persons do not appear to use all the instruments available to them in order to find out whether persons unfamiliar to them are trustworthy or not, such as those provided by modern technology. For example, the Internet provides ample new opportunities to collect and share information about the trustworthiness of individuals or companies with third parties one does not stand in direct contact with, but many of us do not make much use of this.

Much can be gained when constrained maximizers work together in their efforts to distinguish the trustworthy from the untrustworthy. Constrained

maximizers may work together in detecting (un)trustworthy behaviour, as well as sharing information that they have gathered individually. The previous chapters have already mentioned the importance of gossip in this respect. Gossip is a social practice that enables persons to spread information about signs of trustworthiness and in particular of untrustworthiness, albeit a somewhat unreliable one. Constrained maximizers should have no problem taking part in such practices, although they do have reason to increase their reliability. In the same vein, they have reason to support the development of technologies that enable them to track and share information regarding (un)trustworthiness more effectively. As shown by science fiction, instruments are conceivable that increase translucency to such an extent that constrained maximization is much more advantageous than reserved or semi-constrained maximization.

But as there is a limit to how cautious constrained maximizers should be, there may be a limit to the extent to which translucency should be increased. Technologies that increase the translucency of persons may also decrease their privacy. Clear examples are systems that track our whereabouts, the increase in camera surveillance, and the introduction of a general DNA databank. While the introduction of such systems increases the probability of untrustworthy persons being detected, they come at the cost of decreases in privacy. If we are unwilling to accept such trade-offs, which is not implausible given the value that many of us assign to our privacy, there may be a limit to the extent to which translucency can be increased.

This section discussed three types of methods constrained maximizers must employ, and may employ to a greater extent in order to further increase the expected advantage of their own disposition and decrease that of alternative dispositions such as straightforward, reserved, and semi-constrained maximization. By doing so, they may enhance the plausibility of the Translucency Assumption.

10

Conclusions

1 The empirical plausibility of moral contract theory

This investigation concerned the empirical plausibility of moral contract theory. In particular, it concerned the question of whether actual persons have the psychological abilities required for being *contractarian* moral agents.[1] For persons to rationally adopt a contractarian conception of morality, I argued in the introduction, they must first of all be able to grasp the sort of justification that contract theorists give for their conceptions of morality. They must be able to reason about and identify what principles would be the object of unanimous agreement under certain idealised conditions. Besides this cognitive requirement, there is also a conative requirement. Adopting a moral conception involves becoming motivated to comply with its demands, so for agents to be able to adopt a contractarian moral conception they must be able to be moved by its principles.

By proposing their particular conceptions of morality, contract theorists must make certain assumptions regarding these empirical requirements. I have investigated the plausibility of two such assumptions, one regarding the cognitive requirement and another regarding the conative requirement. In this short final chapter I shall summarise my findings. I finish with some practical advice for potential contractarian moral agents.

Let me start by briefly reiterating the two assumptions and why their being correct would support a positive answer. The first assumption, the Practicability Assumption, states that agents are able to find out whether principles for the general regulation of behaviour would be the object of agreement or not. If this assumption is correct, agents are able to understand the justification of the principles of a contractarian moral conception and may

[1] Once again, I use the word 'contractarian' to refer not only to Hobbesian contract theory, but also to the Kantian strain that is often called 'contractualist'.

thus endorse these principles rationally. I have argued in Chapter 2 that for moral contract theorists such as Gauthier and Scanlon, the Practicability Assumption also includes the idea that agents can apply the contractarian justificatory procedure as a test in circumstances of everyday life. If this aspect of the Practicability Assumption is correct, agents can rely on the contract test as a moral guide.

The second assumption relates to the question of whether agents can be motivated to comply with a contractarian conception. As an answer to this question, contract theorists of the Kantian strain such as Rawls and Scanlon have argued that persons already care about morality in such a way that they will be motivated to comply with the demands of their respective moral conceptions. I have not investigated this assumption. Instead, I concentrated on the more controversial idea, defended by Gauthier, that it is in an agent's interest to be motivated to comply with moral demands. A crucial premise of Gauthier's argument is that persons are translucent: that others can see whether they are disposed to comply with moral demands or not. More precisely, he assumes that persons are translucent to such a degree that it is advantageous for agents to be disposed to comply. This Translucency Assumption was the second target of my investigation. As we may expect persons to be motivated by their own interests, if this assumption were correct we may expect that persons can be motivated to comply with moral principles.

The general conclusion of this investigation is that both of these assumptions are empirically plausible. The Practicability Assumption requires that agents can form moral judgments through reasoning that involves perspective-taking, and that they are sufficiently skilled at perspective-taking to find out what persons with other standpoints would agree to. Empirical findings support thinking not just that persons can but in fact do, at least sometimes, form moral judgments through reasoning that involves perspective-taking. And although perspective-taking accuracy is limited such that persons are prone to make mistakes when applying the contract test, we may expect them to be able to overcome these limitations. The key ideas here, about which more in the next section, are preparation and collaboration. The conclusion of Part I is that it is empirically plausible that persons can adopt the contract test as a moral guide.

The Translucency Assumption requires that persons are so skilled at recognising whether others can be trusted to comply with moral norms or not that it is better to be trustworthy. Surely, this does not fit the traditional idea that the minds of others are hidden from us. But this idea turns out to be

mistaken: not only do persons tend to form judgments regarding the trustworthiness of others quickly and effortlessly, these judgments are surprisingly accurate. Indeed, provided they are cautious in choosing with whom to interact, persons are to such a degree translucent that it is more advantageous to be disposed to comply with moral norms than to always do what is in one's own best interest. Whether it is also more advantageous to be trustworthy rather than to be an opportunist who violates norms whenever one has a golden opportunity to do so turns out to be more complicated. There appears to be a small minority who are better off as opportunists than trustworthy persons, notably so-called successful psychopaths. However, empirical findings do not provide reason to think that, for the great majority of agents, being an opportunist is more advantageous than being trustworthy. The conclusion of Part II is that it is empirically plausible that, for by far most agents, internalising certain contractarian moral principles is advantageous.[2]

That these two assumptions are empirically plausible is good news for moral contract theory. Contrary to various concerns and objections, two of its key assumptions about human abilities are in accordance with empirical science. As a theory cannot be plausible as a moral theory if it is not empirically plausible, this conclusion enhances moral contract theory's plausibility as a whole.

2 Some practical advice for contractarian agents

Besides being good news for contract theory, my conclusion that the above two assumptions are empirically plausible is also good news for agents attracted to a contractarian conception of morality. The Practicability Assumption being satisfied would mean that such agents can rely on a contract test as a moral guide. Furthermore, the Translucency Assumption being satisfied would mean that, in so far as principles that satisfy this test overlap with cooperative norms that their interaction partners expect them to adhere to, being disposed to comply with these principles furthers agents' own amoral aims and interests. I suppose such agents would welcome this even if they already care about morality as the contract theorist conceives it, and are as such already motivated to comply with its demands.

---

[2] It bears noting that given that Gauthier's argument concentrates on cooperative norms that one's interaction partners expect one to follow, and as such concerns a rather limited morality, this conclusion is unlikely to generalise to *all* moral principles that a Kantian contract theorist such as Scanlon deems valid.

Being a contractarian agent does require persons to act in ways that may differ from how they tend to do things. Both using the contract test adequately and benefiting from complying with its demands requires persons to follow certain strategies or methods. I will finish by briefly summarising these.

Although persons tend to be inaccurate perspective-takers, there are several methods they may adopt to improve their performance with the contract test. Contractarian agents should embrace these methods and rely on them to improve their moral judgment. I distinguished between three types of methods. First, there are many ways in which persons may become better informed about alternative standpoints (5§2). They may learn about others and their situations by observing them or communicating with them, or through the stories and reports provided by third parties. By becoming more familiar with the situations in which persons may find themselves, agents are less likely to overlook standpoints and relevant objections associated with them.

A second way through which agents can improve their understanding of other standpoints is by relying on the perspective-taking abilities of others (5§3). Due to being differently situated, others may have a better grasp of certain standpoints. An important advantage of this method is that it may correct egocentric biases that our interpretations of other perspectives tend to include. Another advantage is that relevant information about such standpoints available to others is more likely to be taken into account. Besides consulting others regarding their interpretation of particular standpoints, persons may also consult them regarding the acceptability of principles in their entirety. Indeed, when considering difficult, or tricky, cases, it may be advisable to apply the contract test in collaboration with others.

Finally, agents should *prepare* themselves for moral situations by internalising moral principles that satisfy the contract test (5§4). Actual persons who have already acquired a complex system of moral beliefs may use the contract test to reflect on this, shaping and pruning it where necessary. This may be expected to affect the intuitions and affects that often form the starting point of a moral judgment, and can as such lead them to judge in accordance with the contract test even when they have no time to apply it. Moreover, given that moral principles contain information about standpoints and objections, it should also reduce processing costs and time when agents *do* apply the contract test on future occasions.

Given that these methods affect the limitations associated with our capacity for perspective-taking differently, contractarian agents need to rely

on all three of them. And they need to rely on them not just occasionally, but make them part of their practical life. In this way they may ensure to judge and act in accordance with principles that everyone has reason to agree to.

The behavioural implications of the Translucency Assumption for contractarian agents are less straightforward. The agent who is attracted to Gauthier's moral conception because of its promise to make her better off will surely want to act in those ways required by the assumption. But this 'Hobbesian contractarian' must be distinguished from the agent who is attracted to Scanlon's conception because of the centrality that it gives to the values of rational life and mutual respect. Such a 'Kantian contractarian' takes herself to have reason to comply with the principles of Scanlon's conception of morality also when doing so does not further her self-interest, and even when doing so is against her self-interest. However, I presume that such an agent does want to satisfy her interests as best as possible within the constraints provided by her moral ideals. In that case, my recommendations associated with the Translucency Assumption are also relevant for her.

The first method that I discussed in Chapter 9 is that of striving for optimality and against unfairness (§3.1). Contractarian agents are motivated to comply with principles that everyone has reason to accept. They will, however, sometimes find themselves confronted with norms or practices that do not satisfy their standard. Practices may be far from optimal or may involve unfairness. In order to increase their own benefits and reduce the prospect of violating, in particular for persons who would feel unfairly treated by such practices, contractarian agents have reason to support change in the direction of practices that do meet the contract test. Furthermore, contractarian agents should be adamant against being treated unfairly *themselves*. From the perspective of the Translucency Assumption there are two arguments. The first is that persons who are not disposed to respond strongly to unfair offers or unfair treatment will be taken advantage of. The second is that tolerating unfair treatment increases the benefits of being unfair. Note that Kantian contract theorists may pose an additional argument for why contractarian agents should not accept unfair treatment: being treated unfairly violates the respect that others owe them.

The second method required by the Translucency Assumption is that of trusting cautiously (9§3.2). A central property of the contractarian agent as conceived by Gauthier, the constrained maximizer, is that she avoids cooperating with untrustworthy others. Indeed, Gauthier argues that it is not rational to dispose oneself to constrained maximization if one does not at the

same time dispose oneself to exclude others who are not morally committed from profiting from cooperative arrangements. This property not only decreases the probability of being taken advantage of, it also improves the effectiveness of cooperative arrangements. Furthermore, it decreases the expected advantage of not being morally committed.

I have added to this that for the Translucency Assumption to be satisfied, agents must be cautious with whom they interact and what risks they take with them. More precisely, they should typically avoid giving self-benefiting opportunities to persons they have no reason to trust, and instead interact predominantly with others with whom they have a trusting relationship. In addition, I have argued that they should employ the trust-asymmetry principle when forming their expectations: that they take information pertaining to untrustworthiness more seriously than information pertaining to trustworthiness.

While Hobbesian contractarian agents should have no trouble endorsing this second measure, Kantian agents may not go along with it. A Kantian contractarian conception such as Scanlon's may include duties towards persons who have proven themselves to be untrustworthy, and certainly includes such duties towards persons whom one has no reason to trust. There may thus be conflicts between complying with such duties and endorsing this second method.

The final and most obvious method required by the Translucency Assumption is that contractarian agents increase the translucency of their interaction partners (9§3.3). The translucency of one's interaction partners depends on one's social cognitive skills as well as on what one knows about their past behaviour. Both can be improved and should be improved by contractarian agents. As with the Practicability Assumption, collaboration with peers plays a crucial role here. Agents who trust one another should share judgments and information regarding the trustworthiness and untrustworthiness of others. Indeed, it may be a good rule of thumb to only give self-benefiting opportunities to others with whom they are acquainted or about whom they have heard good things from reliable sources.

By embracing the above methods in so far as compatible with the contractarian conception of their preference, persons should be able to become contractarian moral agents. By adopting them they may not only, at least most of the time, be able to act according to principles that everyone has reason to agree to, but also further their own interests while doing so.

Appendix

This appendix shows in detail how we can calculate and compare the expected advantage of constrained and straightforward maximization. I follow Gauthier's analysis, but draw also heavily on Maarten Franssen's (1994), which elaborates on Gauthier's.

1 Comparing constrained and straightforward maximization

Gauthier argues for the rationality of constrained maximization over that of straightforward maximization by arguing that, from the point of view of an agent who can choose between being either a constrained maximizer (CM) or a straightforward maximizer (SM), it is rational to choose to be a CM. As I write in the main text, I follow Gauthier in assuming that the Prisoner's Dilemma (PD) can be used to represent the interactions that this choice concerns. CMs and SMs choose differently in the PD: whereas SMs defect, CMs cooperate if they believe their interaction partner to be a CM as well. Importantly, the agent and her interaction partners are supposed to have either of these two dispositions prior to their interactions; once they meet their dispositions are settled, and they choose as described above.

Table 1: Prisoner's Dilemma, with $e > c > d > f$ and $v > w > x > y$

		Future interaction partners: Cooperate	Future interaction partners: Defect
The agent: Cooperate		c, w mutual cooperation	f, v sucker's payoff, exploitation
The agent: Defect		e, y exploitation, sucker's payoff	d, x noncooperation

Gauthier describes the agent's choice between being a CM or an SM with respect to interactions represented by the PD as an individual decision under risk, and distinguishes four factors that must be taken into account:

- the payoff that the agent expects of the various outcomes. These are represented by the variables e, c, d, f; those of her interaction partners are represented with different variables (v, w, x, y), so to emphasise these do not play a role in the agent's choice;
- the probability p that two CMs recognize each other, which gives the probability $\sqrt{p}$ that a CM is recognised by a CM;
- the probability q that a CM mistakes an SM for a CM;
- the probability r that the agent's interaction partners are CMs. Because Gauthier assumes that, like the agent, the agent's interaction partners are either CMs or SMs, the probability of facing an SM is $1-r$.

Table 2 describes the components of the expected utility for being a CM. The first four cells concern the cases in which the agent meets another CM. The first cell describes the expected utility of mutual recognition, the second and the third that of unilateral recognition, and the fourth that of the case in which neither the agent nor the other CM recognizes one another as such. The last two cells describe when the agent meets an SM. The fifth cell describes the case in which the agent recognises the SM as such, the sixth in which she does not.

Table 2: Expected utility of constrained maximization

CM/rec/rec	CM/rec/nrec	CM/nrec/rec	CM/nrec/nrec	SM/rec/-	SM/nrec/-
crp	$fr\sqrt{p}(1-\sqrt{p})$	$er(1-\sqrt{p})\sqrt{p}$	$dr(1-\sqrt{p})^2$	$d(1-r)(1-q)$	$f(1-r)q$

Taking these components together, the expected utility of being a CM is:

$$crp + fr\sqrt{p}\left(1-\sqrt{p}\right) + er\left(1-\sqrt{p}\right)\sqrt{p} + dr(1-\sqrt{p})^2 + d(1-r)(1-q) + f(1-r)$$

Which comes down to:

(1) $$(c+d-e-f)rp + (e+f-2d)r\sqrt{p} + (d-f)rq + (f-d)q + d$$

Table 3 describes the expected utility of being an SM. The first cell describes that of meeting a CM who does not recognize the agent as an SM, the second that of meeting a CM who does, and the third that of meeting another SM.

Table 3: Expected utility of straightforward maximization

CM/nrec	CM/rec	SM
erq	$dr(1-q)$	$dr(1-r)$

The expected utility of being an SM is:

$$erq + dr(1-q) + d(1-r)$$

Which can be rewritten as:

(2) $$(e-d)rq + d$$

It follows that constrained maximization has a higher expected utility than straightforward maximization if and only if:

(3) $$(e-d)rq < (c+d-e-f)rp + (e+f-2d)r\sqrt{p} + (d-f)rq + (f-d)q$$

2 Proof that $q < \sqrt{p}$

Franssen has shown that that constrained maximization is only more advantageous than straightforward maximization when the probability q that a CM fails to recognise an SM is smaller than the probability $\sqrt{p}$ that a CM recognises a CM.[1] From (3) one can derive what Franssen calls the *generalized Gauthier condition*:

$$(e-d)rq - (d-f)rq - (f-d)q < (c+d-e-f)rp + (e+f-2d)r\sqrt{p}$$

that is

$$q((e+f-2d)r + d - f) < (c+d-e-f)rp + (e+f-2d)r\sqrt{p}$$

[1] A difference with Franssen's analysis and mine is that I use f rather than 0 to represent the sucker's payoff.

so finally

$$(4) \qquad q < \frac{(c+d-e-f)r}{(e+f-2d)r+d-f}p + \frac{(e+f-2d)r}{(e+f-2d)r+d-f}\sqrt{p}$$

Given that per assumption $e > d > f$, the denominator term $((e+f-2d)r + d-f)$ is positive for positive values of f, d and e. Note first that e is a certain value, n, larger than d, and that d is a certain value, m, larger than f. We may thus rewrite d as $f+m$ and e as $f+m+n$. The denominator now becomes $((f+m+n+f-2f-2m)r+f+m-f)$ or $((n-m)r+m)$. Given that $0 \leq r \leq 1$ and therefore $m \geq mr$, this outcome is positive.

It now follows that $q < \sqrt{p}$. Franssen shows this by rewriting the generalized Gauthier condition (4) into:

$$q < \frac{(c+d-e-f)r\sqrt{p}}{(e+f-2d)r+d-f}\sqrt{p} + \frac{(e+f-2d)r}{(e+f-2d)r+d-f}\sqrt{p}$$

that is

$$q < \frac{(c-e)r\sqrt{p} + (d-f)r\sqrt{p} + (e+f-2d)r}{(e+f-2d)r+d-f}\sqrt{p}$$

This final formula implies that $q < \sqrt{p}$. To check this we need to establish that the denominator of the fraction on the right side is larger than its numerator and that the value of this fraction is thus smaller than 1. It then follows that for q to be smaller than the formula on the right, $\sqrt{p}$ must be larger than q.

To see that the denominator of the fraction on the right side is indeed larger than its numerator, note first that the first term of the numerator is negative since $0 < c < e$ is necessary for the PD. Secondly, the second term of the numerator, $(d-f)r\sqrt{p}$, is smaller than the second term of the denominator, $d-f$, for $0 \leq r \leq 1$ and $0 < p < 1$ an $d > f$. Finally, the third term of the numerator is identical to the first term of the denominator. The numerator is thus smaller than the denominator.

It thus follows from (3) that $q < \sqrt{p}$ for whatever values we assign to e, c, d, f and r.

3 Proof that $\sqrt{p}$ must be at least twice as great as q in a symmetric PD

Gauthier (1986, p. 177) holds that there is no reason to expect that, in the interactions an agent will have as either a CM or an SM, the typical gain of exploitation over cooperation is greater or smaller than the gain of cooperation over noncooperation, and no reason to expect this latter gain to be greater or smaller than the typical loss from noncooperation over the sucker's payoff (see 7§2.5). He therefore assumes the differences between payoffs to be equal: $e - c = c - d = d - f$. In the main text I call a PD with this characteristic a *symmetric* PD. Drawing again on Franssen (1994), I shall now show that if we follow Gauthier in representing the structure of an agent's interactions as a symmetric PD, $\sqrt{p}$ must be at least *twice* as large as q for her to be better off as a CM than as an SM.

The crucial property of the symmetric PD is that the difference between the pairs f and d, d and c, c and e, is identical. Call this difference a. We can now represent d as $f + a$, c as $f + 2a$ and e as $f + 3a$, as illustrated by Table 4. Note that that the payoffs of interaction partners are not changed with respect to Table 1; this is because they do not play a role in calculating the expected utility of constrained and straightforward maximization for the agent.

Table 4: Symmetric Prisoner's Dilemma

		Future interaction partners: Cooperate	Future interaction partners: Defect
The agent: Cooperate		$f + 2a, w$	f, v
The agent: Defect		$f + 3a, y$	$f + a, x$

Plugging this in (1), we find the expected utility of constrained maximization to be:

$$\big((f + 2a) + (f + a) - (f + 3a) - f\big)rp + \big((f + 3a) + f - 2(f + a)\big)r\sqrt{p} + \big((f + a) - f\big)rq + \big(f - (f + a)\big)q + (f + a)$$

which can be rewritten as

(1') $\quad ar\sqrt{p} + arq - aq + f + a$

By similar operations, the expected utility of straightforward maximization (2) now becomes:

(2') $2arq + f + a$

The condition under which constrained maximization is more advantageous than straightforward maximization can now be simplified as follows:

$$2arq + f + a < ar\sqrt{p} + arq - aq + f + a$$

that is

(3') $2rq < r\sqrt{p} + rq - q$

From this, one can derive what Franssen calls the *Gauthier condition*, a simplified version of the generalized Gauthier condition (4):

(4') $q < \frac{r}{r+1}\sqrt{p}$

Given that $0 \leq r \leq 1$, the denominator will always be at least twice as large as the numerator. Put differently, the value of the second term is at most $\frac{1}{2}$. It follows that for an agent who is to have interactions with the structure of the symmetric PDs, constrained maximization is only more advantageous than straightforward maximization if $\sqrt{p}$ is at least twice as great as q. While for $r = 1$ this is at the same time sufficient for constrained maximization to be more advantageous, the formula also shows clearly that lower values of r require a greater difference between $\sqrt{p}$ and q.

4 Proofs with regard to 7§4.2

In §4.2 of Chapter 7 I argue that agents should expect that, due to translucency, CMs are more likely to face CMs than SMs are and that, again because of translucency, the interactions of CMs involve on average higher stakes than those of SMs. Each of these assumptions implies, independently, that the expected utility of being a CM increases with respect to that of being an SM. In effect, the degree of translucency required for the expected utility of being a CM to top that of being an SM *decreases*. I shall show this here. For the sake of simplicity I will again assume that the agent's interactions can be

represented by the symmetric PD. I shall for each of these assumptions show that it implies that $\sqrt{p}$ no longer needs to be twice as great as q.

First consider the assumption that CMs are more likely than SMs to face a CM. This assumption can be accounted for by introducing two different probabilities for facing a CM, r_1 for CMs and r_2 for SMs. The expected utility of constrained maximization (1') now becomes:

(1r) $\quad ar_1\sqrt{p} + ar_1q - aq + f + a$

and that of straightforward maximization (2'):

(2r) $\quad 2ar_2q + f + a$

such that constrained maximization is more advantageous than straightforward maximization if and only if:

(3r) $$2r_2q < r_1\sqrt{p} + r_1q - q$$

It is not hard to see that if $r_1 > r_2$, the expected utility of constrained maximization (right hand) has now increased relative to that of straightforward maximization (left hand). It can also easily be shown that this affects how much translucency is required for constrained maximization to be more advantageous than straightforward maximization. With (3r), we can derive an alternative version of the Gauthier condition (4'):

(4r) $$q < \frac{r_1}{2r_2 - r_1 + 1}\sqrt{p}$$

As per assumption $r_1 > r_2$, it is also the case that $r_1 > 2r_2 - r_1$. Contrary to the original Gauthier condition (4'), the value of the second term of this equation may thus increase above 0.5, which implies $\sqrt{p}$ no longer needs to be at least twice as high as q.

Now turn to the assumption that the interactions of CMs involve on average higher stakes than those of SMs. One way to account for this is in the present analysis is by increasing the payoffs that the agent expects as a CM with respect to those she expects as an SM. As it is assumed the agent's interactions can be represented by a symmetric PD, this would have to mean that the gain of cooperation from noncooperation, as well as the gain of exploitation from cooperation and the loss of the sucker's payoff from

noncooperation, increases to a certain (and equal) extent. Call this increase b. Table 5 displays the average payoffs an agent may expect from being a CM; Table 4 displays (as did (2')) what she may expect as an SM.

Table 5: Symmetric Prisoner's Dilemma for CMs

		Future interaction partners: Cooperate	Future interaction partners: Defect
The agent	Cooperate	$f + 2a + 2b, w$	f, v
	Defect	$f + 3a + 3b, y$	$f + a + b, x$

The expected utility of constrained maximization (1') now becomes:

$$\big((f + 2a + 2b) + (f + a + b) - (f + 3a + 3b) - f\big)rp + \big((f + 3a + 3b) + f - 2(f + a + b)\big)r\sqrt{p} + \big((f + a + b) - f\big)rq + \big(f - (f + a + b)\big)q + (f + a + b)$$

which can be rewritten as:

(1b) $$(a + b)\big(r\sqrt{p} + rq - q\big) + f + a + b$$

As b is positive, this means that the expected utility of constrained maximization (1') is increased. Constrained maximization is now more advantageous than straightforward maximization if and only if:

(3b) $$2arq < (a + b)(r\sqrt{p} + rq - q) + b$$

From this we can again derive a version of the Gauthier condition (4'):

$$rq + q < r\sqrt{p} + \frac{b\big(r\sqrt{p} + rq - q + 1\big)}{a}$$

which means that

(4b) $$q < \frac{r}{r + 1}\sqrt{p} + \frac{b\big(r\sqrt{p} + rq - q + 1\big)}{a(r + 1)}$$

Two things should be observed about (4b). First, this formula is identical to the Gauthier condition (4') except for the right hand fraction. Second, this

right hand fraction will have a positive value: a, b are supposed to be positive numbers, as are the probabilities r, $\sqrt{p}$, and q, and $q < 1$. It follows that the formula on the right hand is larger than before (4'). Again, $\sqrt{p}$ no longer needs to be at least twice as high as q.

Both assumptions thus have the same implication: constrained maximization may be more advantageous than straightforward maximization even when the probability $\sqrt{p}$ is not twice as large as the probability q. This means that, everything else being equal, the degree to which people must be translucent for it to be advantageous for agents to be CMs rather than SMs decreases.

If both assumption are justified, that is, if r_1 differs from r_2 and if the payoffs of CMs and SMs differ by b, (3[b]) is transformed into:

(3*) $\quad 2ar_2q < (a + b)(r_1\sqrt{p} + r_1q - q) + b$

Now consider the example given in 7§4.2. The following assumptions were made: $r_1 = 0.9$, $r_2 = 0.75$, $\sqrt{p} = 0.8$, $a = 1$, and $b = 0.25$. Plugging this in (3*) we get:

$$2\times1\times0.75q < 1.25(0.9\times0.8 + 0.9q - q) + 0.25$$

which reduces to

$$q < 0.71$$

SAMENVATTING

Denken over overeenstemming. De empirische plausibiliteit van morele contracttheorie

Inleiding

We hebben allemaal een groot aantal opvattingen over normen en waarden. Ethici zijn geïnteresseerd in de rechtvaardiging van dergelijke opvattingen. Ze ontwikkelen daarom theorieën die uitleggen wat de basis is van onze meer concrete morele opvattingen. Het bekendste voorbeeld van een dergelijke theorie is het utilitarisme, dat stelt dat we zó moeten handelen dat de totale hoeveelheid individueel welzijn zo groot mogelijk is. Een voordeel van zo'n ethische theorie is dat het ook duidelijk kan maken hoe we moeten handelen in moreel lastige situaties. Neem bijvoorbeeld een situatie waarin je moet kiezen tussen de waarheid spreken of iemands gevoelens sparen. Moet je in zo'n geval eerlijk zijn of mag je liegen? Het utilitarisme geeft een duidelijk antwoord: als liegen betere consequenties heeft dan de waarheid spreken, dan moet je liegen.

Contracttheorie, de ethische theorie waar dit proefschrift over gaat, heeft een ander uitgangspunt dan individueel welzijn. Volgens contracttheoretici moeten normen worden gezien als een soort afspraken. Geen afspraken die we daadwerkelijk met elkaar hebben gemaakt, maar afspraken die we met elkaar *zouden* maken als we met elkaar in overleg zouden gaan over de vraag hoe we willen samenleven: morele normen zijn dus principes voor samenleven waar iedereen het mee eens zou kunnen worden. Ook contracttheorie kan gebruikt worden om te bepalen hoe je dient te handelen. Met betrekking tot het eerder genoemde dilemma wordt de relevante vraag: zouden we met elkaar afspreken dat we mogen liegen om elkaars gevoelens te sparen?

Dit proefschrift gaat over de vraag of contracttheorie een plausibele morele theorie is. Preciezer gesteld, het gaat over de vraag of contracttheorie

past bij onze psychologische vermogens. In Hoofdstuk 1 introduceer ik twee aannames die contracttheoretici maken over onze sociaal cognitieve vermogens, onze vermogens om over andere mensen na te denken. De eerste aanname is het idee dat we in staat zijn om te bedenken over welke principes iedereen het eens zou kunnen worden: ik beargumenteer dat contracttheoretici aannemen dat mensen in staat zijn om te 'testen' of handelingen in overeenstemming zijn met dergelijke principes door een zogenaamde 'contracttest' uit te voeren. Het eerste deel van het proefschrift gaat over hoe plausibel deze aanname is in het licht van empirische bevindingen. Het tweede deel gaat over de vraag of het ook in ons belang is om volgens dergelijke principes te leven.

Deel 1: Practicability Assumption

Op het eerste gezicht lijkt het praktisch onmogelijk om te bepalen of iedereen het met een bepaald moreel principe eens zou zijn: zoveel mensen, zoveel meningen. Maar de meeste contracttheoretici denken dat we niet naar iedere individuele mening hoeven te kijken. Zo stelt T.M. Scanlon, wiens contracttheorie in het eerste deel centraal staat, dat we ons moeten richten op abstractere rollen of standpunten die mensen, vanwege hun situatie, hebben ten opzichte van principes. Neem bijvoorbeeld principes die bepalen wanneer mensen elkaar moeten helpen. We kunnen allereerst het standpunt onderscheiden van mensen die door deze principes worden verplicht om anderen te helpen. Daar tegenover staat het standpunt van mensen die door de algemene acceptatie van het principe op hulp van anderen kunnen rekenen. Of een bepaald principe geldig is en dus een principe is op basis waarvan we mogen handelen, is afhankelijk van of het acceptabel is vanuit zulke standpunten.

Dat wil niet zeggen dat er niet ook specifiekere standpunten zijn waar we bij stil moeten staan als we nadenken over wanneer we moeten helpen. Neem het standpunt van mensen in extreme nood. Vanuit dat standpunt zouden we nooit een principe accepteren dat ons toestaat om alleen vrienden en familie te helpen. Aan de andere kant, een principe dat ons verplicht om in de eerste plaats mensen in de allerslechtste posities ter wereld te helpen zal misschien niet acceptabel zijn vanuit het standpunt van onze vrienden en familieleden. Om precies te bepalen wanneer we mensen wel en niet hoeven te helpen zullen we dus een aantal zeer verschillende standpunten moeten bekijken.

Dergelijke voorbeelden maken duidelijk, zo betoog ik in Hoofdstuk 2, dat het toepassen van een contracttest vereist dat we principes bekijken vanuit

meer perspectieven dan alleen onze eigen huidige positie (2§4). Het eerste deel van het proefschrift concentreert zich op de vraag of we dit dermate goed kunnen dat de contracttest een gids voor ons dagelijks leven kan zijn. Om dit te bepalen bespreek ik in Hoofdstuk 3 allereerst of deze manier van denken past bij hoe we daadwerkelijk morele oordelen vormen.

Verschillende moreel psychologen hebben recentelijk beweerd dat het verplaatsen in anderen op z'n hoogst een kleine rol heeft in de vorming van morele oordelen (3§2). Een belangrijk argument voor deze stelling is dat jonge kinderen al morele oordelen vormen terwijl zij nog niet in staat zouden zijn zich in anderen te verplaatsen. Ik beweer dat dit argument niet steekhoudend is (3§2.2). Enerzijds wordt het vermogen van jonge kinderen om de perspectieven van anderen te doorgronden onderschat. Zelfs kinderen onder de twee jaar zijn al in staat om door te hebben dat anderen een verkeerd beeld van de werkelijkheid hebben. Anderzijds *over*schat het argument het moreel oordeelsvermogen van jonge kinderen ook. Empirische studies laten zien dat het morele denken van het jonge kind verre van volwassen is en dat het zich verder ontwikkelt tijdens de jeugd.

Empirisch onderzoek biedt ook bewijs voor de stelling dat het zich kunnen verplaatsen in anderen een belangrijke rol speelt in de vorming van morele oordelen (3§3). Kinderen die beter zijn in het zich verplaatsen in anderen laten ook een hoger niveau van morele ontwikkeling zien. Het lijkt er zelfs op dat ze beter worden in moreel oordelen *doordat* ze beter worden in het zich in anderen verplaatsen. En ook voor volwassenen geldt dat er een positief verband is tussen de neiging om zich in andere perspectieven te verplaatsen en hun moreel denkniveau. Andere studies laten een relatie tussen het verplaatsen in anderen en moreel handelen zien: mensen die het beter of vaker doen zijn ook meer geneigd om anderen te helpen, terwijl asociaal gedrag juist gepaard gaat met een verminderde neiging om zich te verplaatsen in anderen.

Dat we ons soms in anderen verplaatsen om morele oordelen te vormen betekent natuurlijk nog niet dat we er *goed* in zijn. En dat is wel waar contracttheorieën vanuit gaan: het heeft weinig zin om je te laten leiden door een contracttest als je er het grootste deel van de tijd verkeerde conclusies uit trekt. Hoofdstuk 4 onderzoekt daarom of we goed genoeg zijn in ons verplaatsen in anderen om de contracttest adequaat toe te passen.

Wat dit punt betreft geven empirische studies niet een erg optimistisch beeld. We verzaken regelmatig om ons in anderen te verplaatsen, ook al zou het de kwaliteit van onze oordelen kunnen verbeteren (4§2). Het lijkt erop dat we er vaak onterecht vanuit gaan dat anderen hetzelfde perspectief hebben als

wij. En wanneer we ons wel in een ander proberen te verplaatsen, blijven we meestal met tenminste één voet in onze eigen schoenen staan. We hebben de neiging om ervan uit te gaan dat anderen onze overtuigingen, voorkeuren, doelen en zelfs fysieke toestanden zoals dorst delen, en vormen zodoende egocentrische interpretaties van hun perspectief (4§3).

Dat we ons niet of onvoldoende in anderen verplaatsen betekent natuurlijk niet per se dat we ons niet in anderen *kunnen* verplaatsen. In tegendeel, het lijkt erop dat we vaak simpelweg onvoldoende ons best doen (§4.2.2). Maar er zijn daarnaast ook beperkingen aan ons vermogen om ons in anderen te verplaatsen. Zich in anderen verplaatsen kost tijd en aandacht die we soms niet hebben. Het vereist ook informatie over anderen en hun situaties: ik kan moeilijk begrijpen wat een ander nodig heeft als ik niet weet wat hij mist. En zelfs wanneer we voldoende motivatie, tijd, aandacht en informatie hebben, dan nog zijn we geneigd om tot een door ons eigen perspectief gekleurde interpretatie te komen van andere perspectieven.

Betekent dit dat we de contracttest niet goed kunnen toepassen? Het betekent in ieder geval dat we geen natuurtalenten zijn in het gebruik ervan. Maar er lijken manieren te zijn om er beter in te worden, betoog ik in Hoofdstuk 5. Ten eerste kunnen we zorgen dat we beter geïnformeerd zijn over anderen en hun situaties (5§1). Er zijn verschillende manieren om dit te bereiken. Je kunt met mensen in andere situaties communiceren of je kunt hun situaties observeren, hetzij met je eigen ogen, hetzij via de observaties of verhalen van derden. Je kunt zelfs actief ervaring opdoen met nieuwe situaties om zo onbekende standpunten te leren kennen. Ten tweede kunnen we de kans op fouten verkleinen door de contracttest met anderen uit te voeren (5§2). In zoverre dat deze mensen anders zijn dan wij, kunnen we zo ook het egocentrisme van onze eigen interpretaties reduceren. Ten derde kunnen we zorgen dat we morele principes die de contracttest doorstaan internaliseren, zodat we er minder over na hoeven te denken in de toekomst. Deze laatste methode is cruciaal om te zorgen dat we ook in situaties waarin we niet de tijd hebben om de contracttest toe te passen er toch op kunnen rekenen als gids.

Door deze drie methoden goed te gebruiken kunnen mensen de kans op het maken van fouten met de contracttest verkleinen. Daarom concludeer ik dat we kunnen leren de contracttest te gebruiken. De eerste aanname is in overeenstemming met empirische bevindingen.

Deel 2: Translucency Assumption

Waarom zou je je aan morele normen houden? Waarom niet gewoon doen waar je zin in hebt? De contracttheoreticus David Gauthier geeft een verrassend antwoord op deze vraag: het is in je eigenbelang om moreel te zijn en dus heeft iedereen reden om zich aan morele normen te houden. Het tweede deel van het proefschrift gaat over dit idee.

Het is niet moeilijk om te zien dat ik er voordeel van heb als *anderen* zich aan morele normen houden: het betekent immers dat ik er op kan rekenen dat ze mij de waarheid vertellen, dat ze zich aan hun beloften houden en dat ze mij niet zomaar schade zullen toebrengen. Het is ook duidelijk dat anderen er voordeel van hebben als ik me aan de regels houd. Maar dat het mijn belang dient om *zelf* moreel te zijn is verre van duidelijk: het betekent immers dat ik *ook* niet ga liegen, mijn beloften verbreek of op andere manieren schade toebreng aan anderen in die gevallen dat ik daarvan zou profiteren.

Hoofdstuk 6 gaat over het argument van Gauthier. Aan de basis van het argument ligt het idee dat mensen *translucent* of doorschijnend zijn: dat we aan elkaar kunnen zien of we ons aan de normen zullen houden of niet, oftewel, of we *betrouwbaar* zijn of niet. Je kunt maar beter echt betrouwbaar zijn, stelt Gauthier, want anders zullen anderen niet met je willen samenwerken. De centrale vraag is of deze aanname in overeenstemming is met empirische bevindingen.

Een eerste mogelijk bezwaar is dat het idee van doorschijnendheid psychologisch niet aannemelijk is: we kunnen immers niet in elkaars hoofd kijken. Deze tegenwerping houdt geen stand in het licht van empirische studies. Door te kijken naar het gedrag van anderen kunnen we van alles leren over hun mentale toestanden. Dat geldt ook voor betrouwbaarheid: mensen die elkaar observeren vormen binnen een seconde al een oordeel over elkaars betrouwbaarheid en deze oordelen kennen nog een zekere accuraatheid ook (7§2).

Dat wil natuurlijk niet zeggen dat mensen *zo* doorschijnend zijn dat het beter voor hen is om betrouwbaar te zijn dan om morele normen te overtreden wanneer het hen uitkomt. Hier gaat Hoofdstuk 7 over. Studies laten zien dat mensen die vreemden van elkaar zijn verrassend goed zijn in het voorspellen van elkaars betrouwbaarheid (7§2.2-2.4). We mogen verwachten dat deze voorspellingen alleen maar beter worden wanneer het gaat om mensen die eerder contact met elkaar hebben gehad of die informatie over elkaar hebben ontvangen van derden, wat geldt voor de meeste van onze interacties (7§3).

Voeg daaraan toe dat betrouwbare mensen aantrekkelijkere interactiepartners zijn en daarom ook een grotere kans hebben om toegelaten te worden tot bijzonder vruchtbare samenwerkingsverbanden (7§4), en we moeten concluderen dat betrouwbaar zijn verstandiger is dan morele normen overtreden wanneer het je uitkomt.

Maar wat nou als het overtreden van een norm heel voordelig is en de kans dat dit gedetecteerd wordt door anderen wel zeer klein? Het lijkt misschien dat, zolang iemand zich maar voldoende aan de normen houdt zodat hij een goede reputatie heeft als het gaat om betrouwbaarheid, het nemen van zulke 'gouden kansen' verstandiger is dan altijd betrouwbaar zijn. In Hoofdstuk 8 betoog ik dat dit zeer de vraag is. Er is allereerst reden om te denken dat iemand met zo'n opportunistische houding minder betrouwbaar overkomt dan een echt betrouwbare persoon. Om gouden kansen te kunnen pakken moet hij calculerend te werk gaan en kan hij het zich niet permitteren om warme gevoelens te ontwikkelen ten opzichte van anderen. Bovendien heeft hij, vanwege zijn flexibele moraliteit, niet de normale morele sentimenten die het voor betrouwbare mensen gemakkelijker maken elkaar te herkennen. Ten slotte zal hij om gouden kansen te krijgen óf te vermijden dat zijn onbetrouwbare gedrag ontdekt wordt anderen moeten bedriegen en manipuleren. Door dit alles zal zo'n opportunist gemiddeld genomen minder betrouwbaar overkomen dan een daadwerkelijk betrouwbaar persoon en zal hij of zij dus minder gewild zijn als partner in samenwerkingsverbanden.

Dit hoeft geen probleem te zijn voor de opportunist als deze 'kosten' worden gecompenseerd door de voordelen die hij krijgt door gouden kansen. Maar het is onwaarschijnlijk dat dit het geval is. De kans dat een overtreding gedetecteerd zal worden door anderen is in de meeste gevallen zeer moeilijk te voorspellen. Door gebrek aan informatie, denkfouten en beperkte zelfbeheersing loopt een opportunist het risico dat hij normen ook zal overtreden wanneer dit wel gedetecteerd zal worden (8§3). Verder laten empirische studies zien dat vertrouwen inderdaad te voet komt maar te paard gaat (8§4). De kosten van herkend worden als een onbetrouwbare opportunist zijn dus erg hoog.

Daarmee is niet gezegd dat *niemand* beter af zou zijn als een opportunist. We kunnen ons niet alleen mensen voorstellen die bijzonder goed zijn in het bedriegen van anderen, ze lijken ook nog daadwerkelijk te bestaan in de vorm van de psychopaat (8§6). Psychopaten zijn namelijk bijzonder getalenteerd in het bedriegen en manipuleren van anderen. Hoewel veel psychopaten tegelijkertijd zo impulsief zijn dat ze weinig voordeel ondervinden van dit

talent, is er goede reden om te denken dat er ook zogenaamde 'succesvolle psychopaten' bestaan die geen of veel minder last hebben van impulsiviteit. Voor succesvolle psychopaten lijkt opportunistisch te werk te gaan wel degelijk in hun eigenbelang te zijn.

Ik concludeer in Hoofdstuk 9 dat de tweede aanname eveneens in overeenstemming is met empirische bevindingen, mits we deze in één opzicht verzwakken: we zijn dermate doorschijnend dat het niet voor iedereen, maar wel voor de meeste mensen voordelig is om betrouwbaar te zijn.

Conclusie

Beide aannames zijn, met de nodige kwalificaties, empirisch plausibel. Dit is goed nieuws voor contracttheorie: dat belangrijke aannames van de theorie in lijn zijn met empirische bevindingen ondersteunt haar plausibiliteit als morele theorie. Het is ook goed nieuws voor hen die zich aangetrokken voelen tot deze morele theorie. De bevindingen suggereren dat we kunnen zorgen dat ons gedrag in overeenstemming is met principes waar iedereen het redelijkerwijs mee eens kan worden en dat dit in het algemeen ook nog eens in ons eigenbelang is.

Bibliography

Akehurst, L., Bull, R., Vrij, A., & Köhnken, G. (2004). The effects of training professional groups and lay persons to use criteria-based content analysis to detect deception. *Applied Cognitive Psychology, 18*(7), 877–891.

Arriaga, X. B., & Rusbult, C. E. (1998). Standing in my partner's shoes: Partner perspective taking and reactions to accommodative dilemmas. *Personality and Social Psychology Bulletin, 24*(9), 927–948.

Ashford, E., & Mulgan, T. (2012). Contractualism. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2012 Edition)*. Retrieved August 14, 2012, from http://plato.stanford.edu/entries/contractualism/

Babiak, P., & Hare, R. D. (2006). *Snakes in Suits*. New York: Harper.

Babiak, P., Neumann, C. S., & Hare, R. D. (2010). Corporate psychopathy: Talking the walk. *Behavioral Sciences & the Law, 28*(2), 174–193.

Back, M. D., Schmukle, S. C., & Egloff, B. (2010). Why are narcissists so charming at first sight? Decoding the narcissism–popularity link at zero acquaintance. *Journal of Personality and Social Psychology, 98*(1), 132–145.

Baird, J. A., & Astington, J. W. (2004). The role of mental state understanding in the development of moral cognition and moral action. *New Directions for Child and Adolescent Development, 2004*(103), 37–49.

Baron-Cohen, S. (2011). *The Science of Evil: On Empathy and the Origins of Cruelty*. New York: Basic Books.

Barriga, A. Q., Sullivan-Cosetti, M., & Gibbs, J. C. (2009). Moral cognitive correlates of empathy in juvenile delinquents. *Criminal Behaviour and Mental Health, 19*(4), 253–264.

Batson, C. D., & Shaw, L. L. (1991). Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological Inquiry, 2*(2), 107–122.

Batson, C. D., Early, S., & Salvarani, G. (1997a). Perspective taking: Imagining how another feels versus imaging how you would feel. *Personality and Social Psychology Bulletin, 23*(7), 751–758.

Batson, C. D., Polycarpou, M. P., Harmon-Jones, E., Imhoff, H. J., Mitchener, E. C., Bednar, L. L., Klein, T. R., et al. (1997b). Empathy and attitudes: Can feeling for a member of a stigmatized group improve feelings toward the group? *Journal of Personality and Social Psychology, 72*(1), 105–118.

Batson, C. D., Turk, C. L., Shaw, L. L., & Klein, T. R. (1995). Information function of

empathic emotion: Learning that we value the other's welfare. *Journal of Personality and Social Psychology, 68*(2), 300–313.

Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010). The costs and benefits of calculation and moral rules. *Perspectives on Psychological Science, 5*(2), 187–202.

Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy and Public Affairs, 37*(4), 293–329.

Bicchieri, C. (1993). *Rationality and Coordination.* Cambridge: Cambridge University Press.

Bicchieri, C. (2002). Covenants without swords: Group identity, norms and communication in social dilemmas. *Rationality and Society, 14*(2), 192–228.

Bicchieri, C. (2006). *The Grammar Of Society.* Cambridge: Cambridge University Press.

Binder, C. (2012). Walking a day in her shoes. *Spongia: Philosophy Blog EUR.* Retrieved December 21, 2012, from http://philospongia.wordpress.com/2012/03/08/walking-a-day-in-her-shoes

Binmore, K. (1993). Bargaining and morality. In D. Gauthier & R. Sugden (Eds.), *Rationality, Justice and the Social Contract* (pp. 131–156). Hertfortshire: Harvester Wheatsheaf.

Binmore, K. (2005). *Natural Justice.* Oxford: Oxford University Press.

Birch, S. A. J., & Bloom, P. (2003). Children are cursed: An asymmetric bias in mental-state attribution. *Psychological Science, 14*(3), 283–286.

Blair, R. J. R. (2008). Fine cuts of empathy and the amygdala: Dissociable deficits in psychopathy and autism. *The Quarterly Journal of Experimental Psychology, 61*(1), 157–170.

Blakemore, S.-J. (2008). The social brain in adolescence. *Nature Reviews Neuroscience, 9*(4), 267–277.

Bohnet, I., & Frey, B. (1999). The sound of silence in prisoner's dilemma and dictator games. *Journal of Economic Behavior & Organization, 38*(1), 43–57.

Bond, C. F., Jr, & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*(3), 214–234.

Bond, C. F., Jr, & DePaulo, B. M. (2008). Individual differences in judging deception: accuracy and bias. *Psychological Bulletin, 134*(4), 477–492.

Boone, R. T., & Buck, R. (2003). Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation. *Journal of Nonverbal Behavior, 27*(3), 163–182.

Bottom, W. P., Gibson, K., Daniels, S. E., & Murnighan, J. K. (2002). When talk is not cheap: Substantive penance and expressions of intent in rebuilding cooperation. *Organization Science, 13*(5), 497–513.

Braybrooke, D. (1987). Social Contract Theory's Fanciest Flight. *Ethics, 97*(4), 750–764.

Broome, J. (2001). Are intentions reasons? And how should we cope with incommensurable values? In C. W. Morris & A. Ripstein (Eds.), *Practical Rationality and Preference: Essays for David Gauthier*

(pp. 98–120). New York: Cambridge University Press.

Brosig, J. (2002). Identifying cooperative behavior: some experimental results in a prisoner's dilemma game. *Journal of Economic Behavior & Organization, 47*(3), 275–290.

Buchanan, J. (1975). *The Limits of Liberty: Between Anarchy and Leviathan.* Chicago: The University of Chicago Press.

Buchanan, A. (1990). Justice as reciprocity versus subject-centered justice. *Philosophy and Public Affairs, 19*(3), 227–252.

Burt, R. S., & Knez, M. (1996). Trust and third-party gossip. In R. M. Kramer (Ed.), *Trust in Organizations* (pp. 68–89). Thousand Oaks: Sage Publications.

Caruso, E. M., Epley, N., & Bazerman, M. H. (2006). The costs and benefits of undoing egocentric responsibility assessments in groups. *Journal of Personality and Social Psychology, 91*(5), 857–871.

Casebeer, W. D. (2003). Moral cognition and its neural constituents. *Nature Reviews Neuroscience, 4*(10), 840–845.

Castelli, F. (2005). Understanding emotions from standardized facial expressions in autism and normal development. *Autism, 9*(4), 428–449.

Chambers, J. R., Epley, N., Savitsky, K., & Windschitl, P. D. (2008). Knowing too much: Using private knowledge to predict how one is viewed by others. *Psychological Science, 19*(6), 542–548.

Chandler, M. J. (1973). Egocentrism and antisocial behavior: The assessment and training of social perspective-taking skills. *Developmental Psychology, 9*(3), 326–332.

Chiu, C., Hong, Y., & Dweck, C. S. (1997). Lay dispositionism and implicit theories of personality. *Journal of Personality and Social Psychology, 73*(1), 19–30.

Clement, R. W., & Krueger, J. (2000). The primacy of self-referent information in perceptions of social consensus. *British Journal of Social Psychology, 39*(2), 279–299.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development, 9*(4), 377–395.

Coid, J., Yang, M., Ullrich, S., Roberts, A., & Hare, R. D. (2009). Prevalence and correlates of psychopathic traits in the household population of Great Britain. *International Journal of Law and Psychiatry, 32*(2), 65–73.

Coke, J. S., Batson, C. D., & McDavis, K. (1978). Empathic mediation of helping: A two-stage model. *Journal of Personality and Social Psychology, 36*(7), 752–766.

Colby, A., Kohlberg, L., Gibbs, J. C., Lieberman, M., Fischer, K., & Saltzstein, H. D. (1983). A Longitudinal Study of Moral Judgment. *Monographs of the Society for Research in Child Development, 48*, 1–124. Chicago: University of Chicago Press.

Comer, R. J. (2003). *Abnormal Psychology* (Fifth ed.). New York: Worth Publishers.

Copp, D. (1991). Contractarianism and moral skepticism. In P. Vallentyne (Ed.), *Contractarianism and Rational Choice* (pp. 196–228). Cambridge: Cambridge University Press.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (pp. 163–228). New York: Oxford University Press.

Cudd, A. (2012). Contractarianism. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2012 Edition)*. Retrieved December 17, 2012, from http://plato.stanford.edu/entries/contractarianism/

Cvetkovich, G., Siegrist, M., Murray, R., & Tragesser, S. (2002). New information and social trust: Asymmetry and perseverance of attributions about hazard managers. *Risk Analysis, 22*(2), 359–367.

D'Agostino, F., Gaus, G., & Thrasher, J. (2011). Contemporary approaches to the social contract. In E.N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy (Winter 2011 Edition)*. Retrieved December 17, 2012, from http://plato.stanford.edu/archives/win2011/entries/contractarianism-contemporary/

Danielson, P. (1991). Closing the compliance dilemma: How it's rational to be moral in a Lamarckian world. In P. Vallentyne (Ed.), *Contractarianism and Rational Choice* (pp. 291–322). Cambridge: Cambridge University Press.

Darwall, S. (2003). *Contractarianism/Contractualism*. Malden: Blackwell Publishing.

Darwall, S. (2006). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.

Darwin, C. (1899/1998). *The Expression of the Emotions in Man and Animals*. New York: D. Appleton and company. Retrieved December 23, 2012, from http://www.gutenberg.org/cache/epub/1227/pg1227.html

Dasgupta, P. (2000). Trust as a commodity. In D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations* (electronic edition. pp. 49–72). Department of Sociology, University of Oxford.

David, N., Aumann, C., Bewernick, B. H., Santos, N. S., Lehnhardt, F., & Vogeley, K. (2009). Investigation of mentalizing and visuospatial perspective taking for self and other in asperger syndrome. *Journal of Autism and Developmental Disorders, 40*(3), 290–299.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*(1), 113–126.

Davis, M. H., Conklin, L., Smith, A., & Luce, C. (1996). Effect of perspective taking on the cognitive representation of persons: A merging of self and other. *Journal of Personality and Social Psychology, 70*(4), 713–726.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology, 31*(1), 169–193.

Deigh, J. (1995). Empathy and universalizability. *Ethics, 105*(4), 743–763.

Doris, J. (2002). *Lack of Character: Personality and Moral Behavior*. New York: Cambridge University Press.

Doris, J., & Stich, S. (2011). Moral psychology: Empirical approaches. In E.N. Zalta

(Ed.), *The Stanford Encyclopedia (Winter 2011 edition)*. Retrieved August 22, 2012, from http://plato.stanford.edu/entries/moral-psych-emp/

Eisenberg, N. (1986). *Altruistic Emotion, Cognition, and Behavior*. Hillsdale: Lawrence Erlbaum.

Eisenberg, N., Cumberland, A., Guthrie, I. K., Murphy, B. C., & Shepard, S. A. (2005). Age changes in prosocial responding and moral reasoning in adolescence and early adulthood. *Journal of Research on Adolescence, 15*(3), 235–260.

Eisenberg, N., Guthrie, I. K., Cumberland, A., Murphy, B. C., Shepard, S. A., Zhou, Q., & Carlo, G. (2002). Prosocial development in early adulthood: A longitudinal study. *Journal of Personality and Social Psychology, 82*(6), 993–1006.

Eisenberg, N., Miller, P. A., Schaller, M., Fabes, R. A., Fultz, J., Shell, R., & Shea, C. L. (1989). The role of sympathy and altruistic personality traits in helping: A reexamination. *Journal of Personality, 57*(1), 41–67.

Eisenberg, N., Zhou, Q., & Koller, S. (2001). Brazilian adolescents' prosocial moral judgment and behavior: Relations to sympathy, perspective taking, gender-role orientation, and demographic characteristics. *Child Development, 72*(2), 518–534.

Eisenberg, N., & Roth, K. (1980). Development of young children's prosocial moral judgment: A longitudinal follow-up. *Developmental Psychology, 16*(4), 375–376.

Ekman, P. (2003). Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences, 1000*(1), 205–221.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17*(2), 124. American Psychological Association.

Ekman, P., O'Sullivan, M., & Frank, M. G. (1999). A few can catch a liar. *Psychological Science, 10*(3), 263–266. Sage Publications, Inc. on behalf of the Association for Psychological Science.

Elster, J. (1998). Emotions and economic theory. *Journal of Economic Literature, 36*(1), 47–74.

Elster, J. (2007). *Explaining Social Behavior*. New York: Cambridge University Press.

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience, 19*(9), 1508–1519.

Engelmann, D., & Strobel, M. (2000). The false consensus effect disappears if representative information and monetary incentives are given. *Experimental Economics, 3*(3), 241–260.

Epley, N. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology, 40*(6), 760–768.

Epley, N. (2008). Solving the (real) other minds problem. *Social and Personality Psychology Compass, 2*(3), 1455–1474.

Epley, N., & Caruso, E. M. (2004). Egocentric ethics. *Social Justice Research, 17*(2), 171–187.

Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making*, *18*(3), 199–212.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327–339.

Epley, N., Savitsky, K., & Gilovich, T. (2002). Empathy neglect: Reconciling the spotlight effect and the correspondence bias. *Journal of Personality and Social Psychology*, *83*(2), 300–312.

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition*, *26*(2), 143–155.

Fehr, R., Gelfand, M. J., & Nag, M. (2010). The road to forgiveness: A meta-analytic synthesis of its situational and dispositional correlates. *Psychological Bulletin*, *136*(5), 894–914.

Fessler, D. M. T., Arguello, A. P., Mekdara, J. M., & Macias, R. (2003). Disgust sensitivity and meat consumption: a test of an emotivist account of moral vegetarianism. *Appetite*, *41*(1), 31–41.

Fetchenhauer, D., Groothuis, T., & Pradel, J. (2010). Not only states but traits — Humans can identify permanent altruistic dispositions in 20 s. *Evolution and Human Behavior*, *31*(2), 80–86.

Fine, C. (2006). Is the emotional dog wagging its rational tail, or chasing it? *Philosophical Explorations*, *9*(1), 83–98.

Flavell, J. H. (1968). *The Development of Role-Taking and Communication Skills in Children*. New York: John Wiley and Sons.

Ford, G. (2002, April 24). Documentary Special, special episode of *The West Wing*. (W. Couturié, Ed.). New York: NBC.

Fowler, K. A., Lilienfeld, S. O., & Patrick, C. J. (2009). Detecting psychopathy from thin slices of behavior. *Psychological Assessment*, *21*(1), 68–78.

Frank, M. G., & Ekman, P. (1997). The ability to detect deceit generalizes across different types of high-stake lies. *Journal of Personality and Social Psychology*, *72*(6), 1429–1439.

Frank, R. H. (1988). *Passions Within Reason*. New York: W.W. Norton.

Frank, R. H. (2005). Altruists with green beards: Still kicking? *Analyse & Kritik*, *27*, 85–96.

Frank, R. H., Gilovich, T., & Regan, D. T. (1993). The evolution of one-shot cooperation: An experiment. *Ethology and Sociobiology*, *14*(4), 247–256.

Franssen, M. (1994). Constrained maximization reconsidered. An elaboration and critique of Gauthier's modelling of rational cooperation in a single Prisoner's Dilemma. *Synthese*, *101*(2), 249–272.

Freeman, S. (1991). Contractualism, moral motivation, and practical reason. *The*

Journal of Philosophy, *88*(6), 281–303.

Freeman, S. (2006). *Justice and the Social Contract*. Oxford: Oxford University Press.

Freeman, S. (2007). The burdens of public justification: constructivism, contractualism, and publicity. *Politics, Philosophy and Economics*, *6*(1), 5-43.

Freeman, S. (2012). Original position. In E.N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy (Spring 2012 Edition)*. Retrieved August 9, 2012, from http://plato.stanford.edu/archives/spr2012/entries/original-position/

Frei, T. (2009). The redundancy objection, and why scanlon is not a contractualist. *Journal of Political Philosophy*, *17*(1), 47–65.

Friedman, O., & Leslie, A. M. (2007). The conceptual underpinnings of pretense: Pretending is not "behaving-as-if." *Cognition*, *105*(1), 103–124.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*, 459–473.

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, *78*(4), 708–724. 4.

Galinsky, A. D., Magee, J. C., Ena Inesi, M., & Gruenfeld, D. H. (2006). Power and perspectives not taken. *Psychological Science*, *17*(12), 1068–1074.

Gallagher, S., & Hutto, D. (2007). Understanding others through primary interaction and narrative practice. *The Shared Mind: Perspectives on Intersubjectivity*, 17–38.

García-Pérez, R. M., Hobson, R. P., & Lee, A. (2007). Narrative role-taking in autism. *Journal of Autism and Developmental Disorders*, *38*(1), 156–168.

Gauthier, D. (1967). Morality and advantage. *The Philosophical Review*, *76*(4), 460–475.

Gauthier, D. (1969). *The Logic of Leviathan*. Oxford: Oxford University Press.

Gauthier, D. (1975). Reason and maximization. *Canadian Journal of Philosophy*, *4*(3), 411–433.

Gauthier, D. (1977). The social contract as ideology. *Philosophy and Public Affairs*, *6*(2), 130–164.

Gauthier, D. (1986). *Morals by Agreement*. Oxford: Clarendon Press.

Gauthier, D. (1988). Moral artifice. *Canadian Journal of Philosophy*, *18*(2), pp. 385-418.

Gauthier, D. (1990). Justice and natural endowment: Towards a critique of Rawls's ideological framework. In *Moral Dealing. Contract, Ethics, and Reason* (pp. 150–170). New York: Cornell University Press.

Gauthier, D. (1991a). Why contractarianism? In P. Vallentyne (Ed.), *Contractarianism and Rational Choice* (pp. 15–30). Cambridge: Cambridge University Press.

Gauthier, D. (1991b). Rational constraint: Some last words. In P. Vallentyne (Ed.), *Contractarianism and Rational Choice* (pp. 323–330). Cambridge: Cambridge University Press.

Gauthier, D. (1993). Uniting separate persons. In *Rationality, Justice and the Social Contract* (pp. 176–192). Hertfortshire: Harvester Wheatsheaf.

Gauthier, D. (2003). Are we moral debtors? *Philosophy and Phenomenological Research*, *66*(1), 162–168.

Geller, E. (1988). The interplay between linguistic and social-cognitive knowledge in perspective-taking by autistic children. *Communication Disorders Quarterly*, *14*(1), 23–44.

Gibson, K., Bottom, W. P., & Murnighan, J. K. (1999). Once bitten: Defection and reconciliation in a cooperative enterprise. *Business Ethics Quarterly*, *9*(1), 69–85.

Gilovich, T., Medvec, V. H., & Savitsky, K. (2000). The spotlight effect in social judgment: An egocentric bias in estimates of the salience of one's own actions and appearance. *Journal of Personality and Social Psychology*, *78*(2), 211–222.

Gilovich, T., Savitsky, K., & Medvec, V. (1998). The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology*, *75*, 332–346.

Goldman, A. I. (1989). Interpretation psychologized. *Mind and Language*, *4*(3), 161–185.

Goldman, A. I. (2006). *Simulating Minds*. Oxford: Oxford University Press.

Gordon, R. M. (1986). Folk psychology as simulation. *Mind and Language*, *1*(2), 158–171.

Gordon, R. M. (1995). Sympathy, simulation, and the impartial spectator. *Ethics*, *105*(4), 727-742.

Gordon, R. M. (2009). Folk psychology as mental simulation. In E.N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy (Fall 2009 Edition)*. Retrieved December 23, 2012, from http://plato.stanford.edu/archives/fall2009/entries/folkpsych-simulation/

Grant, C. M. (2005). Moral understanding in children with autism. *Autism*, *9*(3), 317–331.

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology. The Neuroscience of Morality: Emotion, Brain Disorders, and Development* (Vol. 3, pp. 35–80). Cambridge, MA: MIT Press.

Greene, J. D., & Baron, J. (2001). Intuitions about declining marginal utility. *Journal of Behavioral Decision Making*, *14*(3), 243–255.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144–1154.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*, 998–1002.

Haidt, J., Björklund, F., & Murphy, S. (2000). Moral Dumbfounding: When Intuition Finds No Reason. *Unpublished Manuscript*, 1–26.

Hall, J. R., & Benning, S. D. (2006). The "successful" psychopath. In *Handbook of Psychopathy*, Handbook of Psychopathy (pp. 459–480). New York: The Guilford Press.

Hampton, J. (1980). Contracts and choices: Does Rawls have a social contract theory? *The Journal of Philosophy*, *77*(6), 315–338.

Hampton, J. (1991). Two faces of contractarianism. In P. Vallentyne (Ed.), *Contractarianism and Rational Choice* (pp. 31–55). Cambridge: Cambridge University Press.

Hampton, J. (1993). Feminist contractarianism. In L. M. Anthony (Ed.), *A Mind of One's Own* (pp. 227–255). Boulder: Westview Press.

Hare, R. D. (1993). *Without Conscience*. New York: The Guilford Press.

Hare, R. M. (1973a). Rawls' Theory of Justice--II. *The Philosophical Quarterly*, *23*(92), 241–252.

Hare, R. M. (1973b). Rawls' Theory of Justice--I. *The Philosophical Quarterly*, *23*(91), 144–155.

Hare, R. M. (1978, October 5). Moral conflicts. *The Tanner Lecture on Human Values*. Given at Utah State University.

Harman, G. (1999). Moral philosophy meets social psychology: virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society*, *99*, 315–331.

Harpending, H. C., & Sobus, J. (1987). Sociopathy as an adaptation. *Ethology and Sociobiology*, *8*, 63–72.

Harris, P. L. (2000). *The Work of the Imagination*. Malden: Blackwell.

Harsanyi, J. C. (1975). Can the maximin principle serve as a basis for morality? A critique of John Rawls's theory. *The American Political Science Review*, *69*(2), 594–606.

Harsanyi, J. C. (1977). *Essays on Ethics, Social Behaviour and Scientific Explanation*. Boston: Reidel.

Hart, S. D., & Dempster, R. (1997). Impulsivity and psychopathy. In C. Webster & M. Jackson (Eds.), *Impulsivity: Theory, Assessment, and Treatment* (pp. 222–232). New York: Guilford Press.

Hartogh, G., den. (1993). The rationality of conditional cooperation. *Erkenntnis*, 38, 405–427.

Helwig, C. C., Zelazo, P. D., & Wilson, M. (2001). Children's judgments of psychological harm in normal and noncanonical situations. *Child Development*, *72*(1), 66–81.

Hobbes, T. (1651/1991). *Leviathan*. (R. Tuck, Ed.). Cambridge: Cambridge University Press.

Hobson, R. P., Lee, A., & Hobson, J. A. (2008). Qualities of symbolic play among children with autism: A social-developmental perspective. *Journal of Autism and Developmental Disorders*, *39*(1), 12–22.

Hodges, J. R. (2008). Making it up and making do: Simulation, imagination, and empathic accuracy. In K. Markman, W. Klein, & J. Suhr (Eds.), *The Handbook of Imagination and Mental Simulation* (pp. 281–294). New York: Psychology Press.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

Holton, R. (2009). *Willing, Wanting, Waiting*. Oxford: Oxford University Press.

Holtzman, N. S. (2011). Facing a psychopath: Detecting the dark triad from

emotionally-neutral faces, using prototypes from the Personality Faceaurus. *Journal of Research in Personality, 45*(6), 648–654.

Houser, D., & Wooders, J. (2006). Reputation in auctions: Theory, and evidence from eBay. *Journal of Economics & Management Strategy, 15*(2), 353–369.

Ickes, W., Stinson, L., Bissonnette, V. L., & Garcia, S. (1990). Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology, 59*(4), 730.

Ittyerah, M., & Mahindra, K. (1990). Moral development and its relation to perspective taking ability. *Psychology & Developing Societies, 2*(2), 203–216.

James, R., & Blair, R. J. R. (1996). Brief report: Morality in the autistic child. *Journal of Autism and Developmental Disorders, 26*(5), 571–579.

Jarrold, C. (2003). A review of research into pretend play in autism. *Autism, 7*(4), 379–390.

Johansson-Stenman, O., Mahmud, M., & Martinsson, P. (2005). Does stake size matter in trust games? *Economics Letters, 88*(3), 365–369.

Jolliffe, D., & Farrington, D. P. (2003). Empathy and offending: A systematic review and meta-analysis. *Aggression and Violent Behavior, 9*(5), 441–476.

Karniol, R. (2003). Egocentrism versus protocentrism: The status of self in social prediction. *Psychological Review, 110*(3), 564–580.

Karpoff, J., & Lott, J., Jr. (1993). The reputational penalty firms bear from committing criminal fraud. *Journal of Law and Economics, 36*(2), 757–802.

Katz, D., Allport, F. H., & Jenness, M. B. (1931). *Students' Attitudes. A Report of the Syracuse University Reaction Study*. Oxford: Craftsman Press.

Kavka, G. S.. (1986). *Hobbesian Moral and Political Theory*. Princeton: Princeton University Press.

Kavka, G. S. (1995). The rationality of rule-following: Hobbes's dispute with the Foole. *Law and Philosophy, 14*, 5–34.

Kawada, C. L. K., Oettingen, G., Gollwitzer, P. M., & Bargh, J. A. (2004). The projection of implicit and explicit goals. *Journal of Personality and Social Psychology, 86*(4), 545–559.

Keysar, B., & Henly, A. S. (2002). Speakers' overestimation of their effectiveness. *Psychological Science, 13*(3), 207–212.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*(1), 32–38.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*(1), 25–41.

Keysers, C., & Gazzola, V. (2006). Towards a unifying neural theory of social cognition. *Progress in Brain Research, 156*, 379–401.

Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition, 119*(2), 197–215.

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- versus integrity-based trust violations. *Journal of Applied Psychology, 89*(1), 104–118.

Klar, Y., & Giladi, E. E. (1999). Are most people happier than their peers, or are they just happy? *Personality and Social Psychology Bulletin, 25*(5), 586–595.

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology, 67*(4), 596–610.

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology, 77*(2), 221–232.

Kruger, J., & Gilovich, T. (1999). "Naive cynicism" in everyday theories of responsibility assessment: On biased assumptions of bias. *Journal of Personality and Social Psychology, 76*(5), 743.

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological Science, 14*(5), 402–408.

Kurzban, R. (2001). The social psychophysics of cooperation: Nonverbal communication in a public goods game. *Journal of Nonverbal Behavior, 25*(4), 241–259.

Kymlicka, W. (1993). The social contract tradition. In P. Singer (Ed.), *A Companion to Ethics* (pp. 186–196). Oxford: Blackwell.

Lane, J. D., Wellman, H. M., Olson, S. L., LaBounty, J., & Kerr, D. C. R. (2010). Theory of mind and emotion understanding predict moral development in early childhood. *British Journal of Developmental Psychology, 28*(4), 871–889.

Ledyard, J. (1995). Public goods: A survey of experimental research. In A. E. Roth & J. Kagel (Eds.), *The Handbook of Experimental Economics* (pp. 111–195). Princeton: Princeton University Press.

Lee, M., & Prentice, N. M. (1988). Interrelations of empathy, cognition, and moral reasoning with dimensions of juvenile delinquency. *Journal of Abnormal Child Psychology, 16*(2), 127–139.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology, 46*(3), 551–556.

Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *American Economic Review, 90*(2), 426–432.

Loewenstein, G., Issacharoff, S., Camerer, C., & Babcock, L. (1993). Self-serving assessments of fairness and pretrial bargaining. *The Journal of Legal Studies, 22*(1), 135–159.

Lott, J., Jr. (1992). An attempt at measuring the total monetary penalty from drug convictions: the importance of an individual's reputation. *The Journal of Legal Studies, 21*(1), 159–187.

Maner, J. K., Luce, C. L., Neuberg, S. L., Cialdini, R. B., Brown, S., & Sagarin, B. J. (2002). The effects of perspective taking on motivations for helping: Still no evidence for altruism. *Personality and Social Psychology Bulletin, 28*(11), 1601–1610.

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition, 112*(3), 367–380.

McIlwain, D., Evans, J., Caldis, E., Cicchini, F., Aronstan, A., Wright, A., & Taylor, A. (2012). Strange moralities: Vicarious emotion and moral emotions in machiavellian and psychopathic personality styles. In R. Langdon & C. Mackenzie (Eds.), *Emotions, Imagination, and Moral Reasoning* (pp. 119–149). New York: Psychology Press.

McLeod, C. (2011). Trust. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*. Retrieved June 4, 2012, from http://plato.stanford.edu/archives/spr2011/entries/trust/

McNaughton, D. (1988). *Moral Vision. An Introduction to Ethics.* Malden: Blackwell Publishing.

Mealey, L., Daood, C., & Krage, M. (1996). Enhanced memory for faces of cheaters. *Ethology and Sociobiology, 17*(2), 119–128.

Messick, D. M., & Sentis, K. P. (1979). Fairness and preference. *Journal of Experimental Social Psychology, 15*(4), 418–434.

Mikhail, J. (2008). Moral cognition and computational theory. In W. Sinnott-Armstrong (Ed.), *Moral Psychology. The Neuroscience of Morality: Emotion, Brain Disorders, and Development* (Vol. 3, pp. 81–92). Cambridge, MA: MIT Press.

Milinski, M., Semmann, D., & Krambeck, H.-J. (2002). Reputation helps solve the tragedy of the commons. *Nature, 415*(6870), 424–426.

Mill, J.S. (1871/2001). *Utilitarianism* (Second Edition). (G. Sher, Ed.). Indianapolis: Hackett Publishing Company.

Moore, C., Jarrold, C., Russell, J., Lumb, A., Sapp, F., & MacCalIum, F. (1995). Conflicting desire and the child's theory of mind. *Cognitive Development, 10*(4), 467–482.

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. D. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences, 108*(7), 2688–2692.

Morris, C. W. (1999). What is this thing called "reputation"? *Business Ethics Quarterly.*

Morris, C. W. (1991). Moral standing and rational-choice contractarianism. In P. Vallentyne (Ed.), *Contractarianism and Rational Choice* (pp. 76–95). Cambridge: Cambridge University Press.

Mullen, B., Atkins, J. L., Champion, D. S., Edwards, C., Hardy, D., Story, J. E., & Vanderklok, M. (1985). The false consensus effect: A meta-analysis of 115 hypothesis tests. *Journal of Experimental Social Psychology, 21*(3), 262–283.

Mullins-Nelson, J., Salekin, R. T., & Leistico, A. (2006). Psychopathy, empathy, and perspective-taking ability in a community sample: Implications for the successful

psychopathy concept. *International Journal of Forensic Mental Health*, *5*(2), 133.

Mullins-Sweatt, S. N., Glover, N. G., Derefinko, K. J., Miller, J. D., & Widiger, T. A. (2010). The search for the successful psychopath. *Journal of Research in Personality*, *44*(4), 554–558.

Myyrya, L., Juujärvi, S., & Pesso, K. (2010). Empathy, perspective taking and personal values as predictors of moral schemas. *Journal of Moral Education*, *39*(2), 213–233.

Nagel, T. (1970). *The Possibility of Altruism*. Princeton: Princeton University Press.

Nelson, A. (1988). Economic rationality and morality. *Philosophy and Public Affairs*, *17*(2), 149–166.

Nelson, J. R., Smith, D. J., & Dodd, J. (1990). The moral reasoning of juvenile delinquents: A meta-analysis. *Journal of Abnormal Child Psychology*, *18*(3), 231–239.

Nesse, R. M. (2001). Natural selection and the capacity for subjective commitment. In R. M. Nesse (Ed.), *Evolution and the Capacity for Commitment*. New York: Russell Sage Foundation.

Nichols, S. (2004). *Sentimental Rules*. New York: Oxford University Press.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, *125*(6), 737–759.

O'Sullivan, M. (2003). The fundamental attribution error in detecting deception: The boy-who-cried-wolf effect. *Personality and Social Psychology Bulletin*, *29*(10), 1316–1327.

Ockenfels, A., & Selten, R. (2000). An experiment on the hypothesis of involuntary truth-signalling in bargaining. *Games and Economic Behavior*, *33*(1), 90–116.

Oda, R. (1997). Biased Face Recognition in the Prisoner's Dilemma Game. *Evolution and Human Behavior*, *18*(5), 309–315.

Oda, R., Yamagata, N., Yabiku, Y., & Matsumoto-Oda, A. (2009). Altruism can be assessed correctly based on impression. *Human Nature*, *20*(3), 331–341.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258.

Oosterbeek, H., Sloof, R., & Van De Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, *7*(2), 171–188.

Oswald, P. A. (1996). The effects of cognitive and affective perspective taking on empathic concern and altruistic helping. *Journal of Social Psychology*, *136*(5), 613–623.

Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.

Parfit, D. (2001). Bombs and coconuts, or rational irrationality. In C. W. Morris & A. Ripstein (Eds.), *Practical Rationality and Preference: Essays for David Gauthier* (pp. 81–97). Cambridge: Cambridge University Press.

Parfit, D. (2011). *On What Matters*. Oxford: Oxford University Press.

Park, H., Levine, T., McCornack, S., & Morrison, K. (2002). How people really detect lies. *Communication Monographs*, *69*(2), 144–157.

Parker, J. G., Rubin, K. H., Erath, S. A., Wojslawowicz, J. C., & Buskirk, A. A. (2006). Peer relationships, child development, and adjustment. In D. Cicchetti & D. J. Cohen (Eds.), *Developmental Psychopathology: Theory and Method* (Vol. 1, pp. 419–493). Hoboken: Wiley & Sons.

Parks, C. D., & Hulbert, L. G. (1995). High and low trusters' responses to fear in a payoff matrix. *Journal of Conflict Resolution*, *39*(4), 718–730.

Piaget, J., & Inhelder, B. (1967). *The Child's Conception of Space*. New York: W. W. Norton & Company.

Pickett, C. L., Gardner, W. L., & Knowles, M. (2004). Getting a cue: The need to belong and enhanced sensitivity to social cues. *Personality and Social Psychology Bulletin*, *30*(9), 1095–1107.

Pizarro, D. A. (2000). Nothing more than feelings? The role of emotions in moral judgment. *Journal for the Theory of Social Behaviour*, *30*(4), 355–375.

Pizarro, D. A., & Bloom, P. (2003). The intelligence of the moral intuitions: A comment on Haidt (2001). *Psychological Review*, *110*(1), 193–196.

Poortinga, W., & Pidgeon, N. (2004). Trust, the asymmetry principle, and the role of prior beliefs. *Risk Analysis*, *24*(6), 1475–1486.

Porter, S., & Brinke, L., ten. (2008). Reading between the lies. *Psychological Science*, *19*(5), 508–514.

Porter, S., & Brinke, L., ten. (2011). The truth about lies: What works in detecting high-stakes deception? *Legal and criminological Psychology*, *15*(1), 57–75.

Porter, S., Brinke, L., ten, & Wilson, K. (2009). Crime profiles and conditional release performance of psychopathic and non-psychopathic sexual offenders. *Legal and criminological Psychology*, *14*(1), 109–118.

Porter, S., Brinke, L., ten, Baker, A., & Wallace, B. (2011). Would I lie to you? "leakage" in deceptive facial expressions relates to psychopathy and emotional intelligence. *Personality and Individual Differences*, *51*(2), 133–137.

Prinz, J. J. (2007). *The Emotional Construction of Morals*. Oxford: Oxford University Press.

Prinz, J. J. (2011). Is Empathy Necessary for Morality? In A. Coplan & P. Goldie (Eds.), *Empathy: Philosophical and Psychological Perspectives* (pp. 211-229). Oxford: Oxford University Press.

Rawls, J. (1974). The independence of moral theory. *Proceedings and Addresses of the American Philosophical Association*, *48*, 5–22.

Rawls, J. (1999). *A Theory of Justice* (Revised Edition). Cambridge, MA: The Belknap Press of Harvard University Press.

Resnick, P., Kuwabara, K., Zeckhauser, R., & Friedman, E. (2000). Reputation systems. *Communications of the ACM*, *43*(12), 45–48.

Rest, J. R. (1999). *Postconventional Moral Thinking: A Neo-Kohlbergian Approach*. Mahwah: Lawrence Erlbaum.

Rest, J. R., Narvaez, D., Thoma, S., & Bebeau, M. (2000). A Neo-Kohlbergian approach to morality research. *Journal of Moral Education*, *29*(4), 381–395.

Richardson, D. R., Green, L. R., & Lago, T. (1998). The relationship between perspective-taking and nonaggressive responding in the face of an attack. *Journal of Personality*, *66*(2), 235–256.

Roge, B., & Mullet, E. (2011). Blame and forgiveness judgements among children, adolescents and adults with autism. *Autism*, *16*(6), 702–712.

Ross, D., & Dumouchel, P. (2004). Emotions as strategic signals. *Rationality and Society*, *16*(3), 251–286.

Ross, L., Greene, D., & House, P. (1977). The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279–301.

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology*, *37*(3), 322.

Rozin, P., Markwith, M., & Stoess, C. (1997). Moralization and becoming a vegetarian: The transformation of preferences into values and the recruitment of disgust. *Psychological Science*, *8*(2), 67–73.

Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, *7*(1), 58–92.

Sauer, H. (2012). Educated intuitions. Automaticity and rationality in moral judgement. *Philosophical Explorations*, *15*(3), 255–275.

Savitsky, K., Epley, N., & Gilovich, T. (2001). Do others judge us as harshly as we think? Overestimating the impact of our failures, shortcomings, and mishaps. *Journal of Personality and Social Psychology*, *81*(1), 44–56.

Savitsky, K., Van Boven, L., Epley, N., & Wight, W. M. (2005). The unpacking effect in allocations of responsibility for group tasks. *Journal of Experimental Social Psychology*, *41*(5), 447–457.

Sayre-McCord, G. (1991). Deception and reasons to be moral. In *Contractarianism and Rational Choice* (pp. 181–195). Cambridge: Cambridge University Press.

Scanlon, T. M. (1982). Contractualism and utilitarianism. In A. Sen & B. Williams (Eds.), *Utilitarianism and beyond* (pp. 103–128). Cambridge: Cambridge University Press.

Scanlon, T. M. (1998). *What we owe to each other*. Cambridge, MA: The Belknap Press of Harvard University Press.

Scanlon, T. M. (2003). Reply to Gauthier and Gibbard. *Philosophy and Phenomenological Research*, *66*(1), 176–189.

Scanlon, T. M. (2011). How I am not a Kantian. In D. Parfit, *On What Matters* (Vol. II, pp. 116–141). Oxford: Oxford University Press.

Schug, J., Matsumoto, D., Horita, Y., Yamagishi, T., & Bonnet, K. (2010). Emotional expressivity as a signal of cooperation. *Evolution and Human Behavior*, *31*(2), 87–94.

Schweitzer, M. E., Hershey, J., & Bradlow, E. (2006). Promises and lies: Restoring

violated trust. *Organizational Behavior and Human Decision Processes*, *101*(1), 1–19.

Selman, R. L. (1971). The relation of role taking to the development of moral judgment in children. *Child Development*, *42*(1), 79–91.

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in asperger syndrome. *Science*, *325*, 883–885.

Shamay-Tsoory, S. G., Tomer, R., Yaniv, S., & Aharon-Peretz, J. (2002). Empathy deficits in asperger syndrome: A cognitive profile. *Neurocase*, *8*(3), 245–252.

Shapiro, D. L. (1991). The effects of explanations on negative reactions to deceit. *Administrative Science Quarterly*, *36*(4), 614–630.

Shapiro, D. L., Buttner, E. H., & Barry, B. (1994). Explanations: What factors enhance their perceived adequacy? *Organizational Behavior and Human Decision Processes*, *58*(3), 346–368.

Shaver, R. (2010). Egoism. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Winter 2010 Edition)*. Retrieved May 29, 2012, from http://plato.stanford.edu/archives/win2010/entries/egoism/

Siegler, R. S., DeLoache, J. S., & Eisenberg, N. (2011). *How Children Develop* (Third edition). New York: Worth Publishers.

Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

Slovic, P. (1993). Perceived risk, trust, and democracy. *Risk Analysis*, *13*(6), 675–682.

Slovic, P. (1999). Trust, emotion, sex, politics, and science: Surveying the risk-assessment battlefield. *Risk Analysis*, *19*(4), 689–701.

Slovic, P., & Peters, E. (2006). Risk perception and affect. *Current Directions in Psychological Science*, *15*(6), 322–325.

Smetana, J. G. (1985). Preschool children's conceptions of transgressions: Effects of varying moral and conventional domain-related attributes. *Developmental Psychology*, *21*(1), 18–29.

Smetana, J. G., & Braeges, J. L. (1990). The development of toddler's moral and conventional judgments. *Merrill-Palmer Quarterly*, *36*(3), 329–346.

Sociopath, A. (2010, May 31). Comment to: Do Sociopaths know they are sociopaths? *SociopathWorld*. Retrieved December 23, 2012, from http://www.sociopathworld.com/2008/11/do-sociopaths-know-they-are-sociopaths.html

Southwood, N. (2010). *Contractualism and the Foundation of Morality*. New York: Oxford University Press.

Sperber, D., Clement, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Deirdre, W. (2010). Epistemic vigilance. *Mind and Language*, *25*(4), 359–393.

Spiekermann, K. P. (2007). Translucency, assortation, and information pooling: how groups solve social dilemmas. *Politics, Philosophy and Economics*, *6*(3), 285–306.

Stams, G. J., Brugman, D., Deković, M., Rosmalen, L., Laan, P., & Gibbs, J. C. (2006). The moral judgment of juvenile delinquents: A meta-analysis. *Journal of Abnormal Child Psychology*, *34*(5), 692–708.

Standifird, S. S. (2001). Reputation and e-commerce: eBay auctions and the asymmetrical impact of positive and negative ratings. *Journal of Management*, *27*(3), 279–295.

Stinson, L., & Ickes, W. (1992). Empathic accuracy in the interactions of male friends versus male strangers. *Journal of Personality and Social Psychology*, *62*(5), 787.

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science*, *21*(3), 349–354.

Struthers, C. W., Eaton, J., Santelli, A. G., Uchiyama, M., & Shirvani, N. (2008). The effects of attributions of intent and apology on forgiveness: When saying sorry may not help the story. *Journal of Experimental Social Psychology*, *44*(4), 983–992.

Sugden, R. (1993). Rationality and impartiality: Is the contractarian enterprise possible? In D. Gauthier & R. Sugden (Eds.), *Rationality, Justice, and the Social Contract* (pp. 157–175). Exeter: Harvester Wheatsheaf.

Takagishi, H., Kameshima, S., Schug, J., Koizumi, M., & Yamagishi, T. (2010). Theory of mind enhances preference for fairness. *Journal of Experimental Child Psychology*, *105*(1–2), 130–137.

Tetlock, P. E. (2003). Thinking the unthinkable: sacred values and taboo cognitions. *Trends in Cognitive Sciences*, 7(7), 320–324.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, 78(5), 853–870.

Thompson, M. (2008). *Life and Action: Elementary Structures of Practice and Practical Thought*. Cambridge, MA: Harvard University Press.

Tomasello, M. (2009). *Why We Cooperate*. Cambridge: MIT Press.

Underwood, B., & Moore, B. (1982). Perspective-taking and altruism. *Psychological Bulletin*, *91*(1), 143–173.

Unger, L. S., & Thumuluri, L. K. (1997). Trait empathy and continuous helping: The case of voluntarism. *Journal of Social Behavior & Personality*, *12*(3), 785–800.

Vallentyne, P. (1991a). Gauthier's three projects. In P. Vallentyne (Ed.), *Contractarianism and Rational Choice* (pp. 1–11). Cambridge: Cambridge University Press.

Vallentyne, P. (1991b). Contractarianism and the assumption of mutual unconcern. In P. Vallentyne (Ed.), *Contractarianism and Rational Choice* (pp. 71–75). Cambridge: Cambridge University Press.

Van Boven, L., & Loewenstein, G. (2003). Social projection of transient drive states. *Personality and Social Psychology Bulletin*, *29*(9), 1159–1168.

Van Boven, L., & Loewenstein, G. (2005). Cross-situational projection. In M. D. Alicke, D. Dunning, & J. I. Krueger (Eds.), *The Self in Social Judgment* (pp. 43–64). New York: Psychology Press.

Van Boven, L., Dunning, D., & Loewenstein, G. (2000). Egocentric empathy gaps between owners and buyers: Misperceptions of the endowment effect. *Journal of*

Personality and Social Psychology, 79(1), 66–76.

Van Boven, L., Loewenstein, G., & Dunning, D. (2003). Mispredicting the endowment effect:: Underestimation of owners' selling prices by buyer's agents. *Journal of Economic Behavior & Organization, 51*(3), 351–365.

Van Boven, L., Loewenstein, G., & Dunning, D. (2005). The illusion of courage in social predictions: Underestimating the impact of fear of embarrassment on other people. *Organizational Behavior and Human Decision Processes, 96*(2), 130–141.

Vanneste, S., Verplaetse, J., van Hiel, A., & Braeckman, J. (2007). Attention bias toward noncooperative people. A dot probe classification study in cheating detection. *Evolution and Human Behavior, 28*(4), 272–276.

Verplaetse, J., Vanneste, S., & Braeckman, J. (2007). You can judge a book by its cover: the sequel.A kernel of truth in predictive cheating detection. *Evolution and Human Behavior, 28*(4), 260–271.

Vignemont, F., de, & Frith, U. (2008). Autism, morality, and empathy. In W. Sinnott-Armstrong (Ed.), *Moral Psychology. The Neuroscience of Morality: Emotion, Brain Disorders and Development* (Vol. 3, pp. 273–280). Cambridge, MA: MIT Press.

Vorauer, J. D., & Claude, S. D. (1998). Perceived versus actual transparency of goals in negotiation. *Personality and Social Psychology Bulletin, 24*(4), 371–385.

Vorauer, J. D., & Ross, M. (1999). Self-awareness and feeling transparent: Failing to suppress one's self. *Journal of Experimental Social Psychology, 35*(5), 415–440.

Vrij, A. (2004). Why professionals fail to catch liars and how they can improve. *Legal and criminological Psychology, 9*, 159–181.

Vrij, A. (2005). Criteria-based content analysis: A qualitative review of the first 37 studies. *Psychology, Public Policy, and Law, 11*(1), 3–41.

Vrij, A., Edward, K., & Bull, R. (2001). Stereotypical verbal and nonverbal responses while deceiving others. *Personality and Social Psychology Bulletin, 27*(7), 899–909.

Vrij, A., Granhag, P. A., & Porter, S. (2011a). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest, 11*(3), 89–121.

Vrij, A., Granhag, P. A., Mann, S., & Leal, S. (2011b). Outsmarting the liars: Toward a cognitive lie detection approach. *Current Directions in Psychological Science, 20*(1), 28–32.

Vrij, A., Mann, S. A., Fisher, R. P., Leal, S., Milne, R., & Bull, R. (2008). Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior, 32*(3), 253–265.

Vrij, A., Mann, S., Kristen, S., & Fisher, R. P. (2007). Cues to deception and ability to detect lies as a function of police interview styles. *Law and Human Behavior, 31*(5), 499–518.

Wallace, R. J. (2008). Practical reason. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Summer 2009 Edition)*. Retrieved February 10, 2012, from http://plato.stanford.edu/archives/sum2009/entries/practical-reason

Watson, G. (1998). Some considerations in favor of contractualism. In C. Morris & J. L. Coleman (Eds.), *Rational Commitment of Social Justice: Essays for Gregory Kavka*

(pp. 168–185). Cambridge: Cambridge University Press.

Waytz, A., & Epley, N. (2012). Social connection enables dehumanization. *Journal of Experimental Social Psychology, 48*(1), 70–76.

Waytz, A., Klein, N., & Epley, N. (forthcoming). Imagining other minds: Hair triggered but not hare brained. In M. Taylor (Ed.), *The Development of Imagination*. New York: Oxford University Press.

White, M. P., Pahl, S., Buehner, M., & Haye, A. (2003). Trust in risky messages: The role of prior attitudes. *Risk Analysis, 23*(4), 717–726.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592–598.

Yamagishi, T., Tanida, S., Mashima, R., Shimoma, E., & Kanazawa, S. (2003). You can judge a book by its cover. Evidence that cheaters may look different from cooperators. *Evolution and Human Behavior, 24*(4), 290–301.

Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E., & Chapman, M. (1992). Development of concern for others. *Developmental Psychology, 28*(1), 126–136.

Zaitchik, D. (1991). Is only seeing really believing?: Sources of the true belief in the false belief task. *Cognitive Development, 6*(1), 91–103.

Zalla, T., & Leboyer, M. (2011). Judgment of intentionality and moral evaluation in individuals with high functioning autism. *Review of Philosophy and Psychology, 2*, 681–698.

Zalla, T., Sav, A.-M., Stopin, A., Ahade, S., & Leboyer, M. (2009). Faux pas detection and intentional action in asperger syndrome. A replication on a french sample. *Journal of Autism and Developmental Disorders, 39*(2), 373–382.

Zelazo, P. D., Helwig, C. C., & Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. *Child Development, 67*(5), 2478–2492.