Institutional Self-Deception

A thesis submitted in fulfilment of the
requirements for the degree of
DOCTOR OF PHILOSOPHY
of
RHODES UNIVERSITY

By GENEVIEVE JACOT-GUILLARMOD

February 2022

Abstract:

There are many examples of institutions which have made false claims, or performed certain acts, that have had, to varying degrees, a negative impact on their societies. For example, many corporations go to great lengths to present themselves as being environmentally friendly when in fact they are not. Many corporations have also been forced to recall dangerous products which they at one time or another insisted were safe. Research teams have misled participants with regard to what they can expect from their participation in studies, with grave consequences. Governments throughout the world are mired in corruption, and yet deny that this is so. One possible explanation is that in such situations these institutions are simply lying. However, another possible explanation is that these institutions are self-deceived, or lying to themselves.

Recently it has been suggested that self-deception is something that affects certain groups as well as individuals. Given that institutions can wield a great deal of political, social and economic power, if institutions are capable of self-deception there is room for things to go awry on a very large scale with potentially dire consequences. Yet the explanations currently on offer for group-level self-deception appear to amount to instances of individual self-deception (either to certain key individual members of those groups being self-deceived, or to all or most members of a group sharing the same self-deceptive belief), and as such I do not regard the explanations currently on offer as satisfactory. I propose that there are certain situations in which we ought to see institutions themselves as self-deceived or lying to themselves.

While the terms 'self-deception' and 'lying to oneself' are often used interchangeably, I differentiate between the two and argue that both institutional self-deception and an institution lying to itself are institution-level phenomena, and do not rely on any individual within the institution being self-deceived or lying to themselves. That this is so is of

relevance to our attributions of accountability, and makes changes to institutional structure and procedures the focus of concern when it comes to preventing an institution succumbing to self-deception or lying to itself.

Contents:

Introduction …p1

Chapter 1: What are Institutions, and Why Should We be Interested in Them? …p6

Chapter 2: Making Sense of ‘Group Mind’ Talk When it Comes to Institutions …p22

Chapter 3: Institutions can be Minded and Moral Agents …p43

Chapter 4: Institutions can be Liars …p69

Chapter 5: Institutions can be Self-Deceived and Lie to Themselves …p83

Chapter 6: Moral Implications and Possible Solutions …p99

Conclusion …p118

References …p120

Acknowledgements

I would like to express my sincere thanks to my supervisor, Ward Jones, for his extremely helpful and constructive feedback, recommendations and advice throughout. I would also like to extend my thanks to the Rhodes University Philosophy Department for their helpful comments and suggestions on earlier drafts. A special thanks to Tess Dewhurst for her detailed reading of a final draft and her helpful recommendations for improvements.
Last but not least, I would also like to thank my family and friends for their endless support and encouragement.

Introduction

There are many examples of institutions that have claimed things which are not the case, or that have put forward certain views and yet behaved contrary to those views, which have had, to varying degrees, a negative impact on their societies. For example, many corporations have gone to great lengths to present themselves as being environmentally friendly when in fact they are not, misleading the public.[1] Many corporations have also been forced to recall dangerous products which they at one time or another insisted were safe.[2] Educational institutions have misled the public about their academic track record.[3] Research teams have often misled participants with regard to what they can expect from their participation in studies, with grave consequences.[4] Governments throughout the world are often mired in corruption, and yet deny that this is so.[5] One possible explanation is that in such situations these institutions are simply lying when they put forward such views, present themselves in such ways, or make such denials (see Lackey, 2018). However, another possible explanation for putting forward false claims along the lines of those above is that these institutions are self-deceived.

In recent years it has been suggested that self-deception is something that can affect groups such as institutions as well as individuals. However, the explanations currently on offer for group level self-deception appear to amount either to certain key individual members of those groups being self-deceived (Trivers, 2000), or to all or most members of a group sharing the same self-deceptive belief (Ariely, 2012; Deweese-Boyd, *forthcoming*). It is certainly true that group dynamics can play a role in the strength of self-deceptive beliefs of individual group members, and make self-deception more difficult to escape (more on this in Chapter 5). Yet given the prevalence of the view that "any talk of group agents is either metaphorical or wrong" (List & Pettit 2011, p3), the explanations currently on offer leave one wondering whether or not the phenomenon of group or institutional self-deception is

[1] E.g., Kauai Coffee's somewhat misleadingly labelled 100% compostable coffee pods have only been shown to decompose at industrial facilities, and not in compost heaps at consumers' homes.
[2] E.g., Ford Motor Company's recall of the Pinto.
[3] E.g., the University of North Carolina claimed to be able to maintain high academic standards while simultaneously running a successful top sport programme, yet in actual fact it was giving credits for fake courses to many students that engaged in the demanding sports programmes.
[4] E.g., in the San Antonio Contraceptive Study, a number of Mexican-American women were given placebos rather than actual contraceptive pills without being made aware that this was a possibility – many of them had unplanned pregnancies as a result.
[5] E.g., evidenced by the long-standing corruption charges against a number of South African politicians (both former and current).

genuinely distinguishable from that of individual self-deception. Ultimately, I do not regard the explanations currently on offer as satisfactory, and my aim will be to show that there are certain situations in which we ought to see institutions themselves as self-deceived or lying to themselves. These are institution-level phenomena, and do not rely on any individual within the institution being self-deceived or lying to themselves. That this is so will be of relevance to our attributions of individual accountability.

An important methodological consideration in this thesis will involve offering an interpretation of our practice of attributing mental states and behaviour to institutions as a whole. In doing so I will be interested in taking this practice seriously, and I will be interested in assessing the arguments of others who view this practice as more than merely metaphorical. I hope thereby to tackle the question posed by Deborah Tollefsen when she asks, "[w]hat are we to make of this practice?" (2015, p2). I will also look at our practices of interaction with institutions, and what we can infer from these. I follow Peter Strawson in arguing that our practices can reveal certain attitudes which we have towards groups like institutions, as for Strawson our "practices […] really are expressions of our moral attitudes and not merely devices we calculatingly employ for regulative purposes" (Strawson, 1962, p27).

I will begin my argument in Chapter 1 with a general discussion of the nature of groups, and provide motivation for my selected focus on institutions (governments, corporations, universities etc.) rather than other kinds of groups as institutions often wield a great deal of social, political and economic power. Given that self-deception is at least potentially dangerous when individuals succumb to it, if institutions can be self-deceived then there is arguably room for things to go awry on a much greater scale as institutions tend to wield far more widespread power and influence than individuals.

In order for institutions to be said to lie, lie to themselves, or be self-deceived, they must be minded and moral agents. Lying itself not only requires the ability to have beliefs and intentions, it is also a moral notion. Self-deception is also at least generally regarded as a morally problematic phenomenon, and commonly understood to revolve around beliefs and false beliefs. If institutions do not have beliefs of their own, we may have to reassess an institution's ability to lie. We will also have to set aside talk of 'institutional' self-deception and instead understand what may appear to be institution-level self-deception in terms of individual self-deception or perhaps interpersonal deception. Consequently, in Chapters 2 and 3, I will argue that institutions can rightly be seen as intentional agents with beliefs and desires – and so self-deceptive beliefs – of their own.

In establishing the above, I hope to also highlight the degree to which institutions are often capable of wielding an enormous amount of power over our lives, as well as the fact that they are capable of behaving in an unethical manner. In Chapter 4 I look more carefully at the claim that institutions themselves can lie, and examine the harms that can be brought about by institutional lies. As I indicate above, an alternative explanation and, as I will argue, better explanation for some of the unethical behaviour we witness is that an institution is not lying, but rather lying to themselves or self-deceived.

In Chapter 5 I will argue that institutions are capable of lying to themselves and of being self-deceived, and that such states can, like lying, be just as harmful – indeed, the very same harms may be brought about. Although it can often be difficult to determine beyond all doubt whether or not an institution is merely lying, or whether it is in fact lying to itself or self-deceived, I will attempt to clearly distinguish between these three phenomena with a view to determining how can assess whether or not an institution is lying, lying to itself or self-deceived, and we ought to approach instances of an institution lying to itself or institutional self-deception.

Historically, formulating a comprehensive philosophical account of self-deception has proven challenging. The notion of self-deception seems, at least initially, to be paradoxical. Attempts to model self-deception on ordinary cases of interpersonal deception (see Demos, 1960; Sartre, 1943) require that the self-deceived take on the role of both the deceiver and the deceived, something which looks incoherent. This is because someone must believe that something is the case (i.e., that *p* is true) in order to be able to lie about it, yet at the same time they must fail to think that such and such is the case in order to be taken in by the lie (i.e., come to believe that *p* is false). The apparent incoherence of someone lying to herself has prompted a number of philosophers to adopt what has been dubbed a 'deflationary' approach to the phenomenon. While it is generally accepted that individual self-deception minimally involves acquiring or retaining beliefs in the face of strong evidence to the contrary (where the retention of such beliefs is motivated by one's desires or other emotions [Bortolotti & Mameli, 2012]), such accounts do not model self-deception on interpersonal deception. Instead, they suggest that self-deception is generated through an unintentional process (that lies beyond one's conscious awareness). Alfred Mele has provided an account in this vein which successfully avoids the paradoxes attendant on self-deception without the need for any of what he dubs "mental exotica" in order to make sense of the phenomenon (Mele, 2001, p4). One way of seeing whether or not institutions are capable of self-deception,

then, is to see whether or not they can meet Mele's four minimal conditions for self-deception. In Chapter 5 I shall argue that they can.

However, institutions also seem to offer a fertile setting for a very literal take on self-deception (something that, as noted above, causes trouble for an account of individual self-deception). Given that institutions are comprised of persons, institutions can lie to themselves through members' lying to each other – indeed, that persons themselves are not comprised of sub-agents is what causes trouble for this more literal interpretation of self-deception. It may seem at first as if the situation I am describing should rather be categorised as cases of interpersonal deception, however I will argue that if the beliefs involved are those of the institution (regardless of whether or not they are also beliefs of the relevant members of that institution), then the institution can rightly be regarded as lying to itself. With this in mind I will argue in Chapter 5 that not only are institutions capable of self-deception understood in deflationary terms, they are also capable of self-deception as traditionally conceived, or what I will term "lying to oneself".

It has been proposed that not all self-deceptive beliefs are bad (that self-deception regarding certain aspects of ourselves or our lives may actually be necessary in order for us to function in the world, and indeed could have an evolutionary basis [Taylor, 1989; Trivers 2000]). Yet it is also the case that self-deception, or lying to oneself, can often make us unaware of unpleasant truths, or our own moral shortcomings, and so can prevent us from addressing these problems and defects. It may also be the case that, as a result of self-deception, we can end up harming others or ourselves. Hence self-deception is largely viewed as a highly problematic phenomenon, and something that one would do well to avoid.

Acknowledging that institutions are capable of self-deception will have implications for how we think about the responsibility of institutions, and the ways in which responsibility ought to be distributed within such groups. Indeed, a closer examination of how self-deception is instantiated within institutions could elucidate to what degree institutions as a whole should be held accountable for actions that fall under their sway.

Nevertheless, if one applies the most prevalent deflationary account of self-deception to institutions, one may worry that institutions cannot be held responsible for their self-deceptive beliefs and any behaviour that stems from these beliefs, because such accounts see self-deception as something which is unintentional. In Chapter 6 I outline the impact of this on issues surrounding the responsibility of institutions, and argue that institutions can still be held responsible for self-deception and actions that stem from it. Institutions that perform harmful acts as a result of self-deception could still rightly be expected to make public

apologies, pay fines, make reparations, or make structural changes as the case may be. I also will argue that when institutions are lying to themselves, an absence of responsibility does not follow.

Finally, it is important to accurately identify the source of an institution's unethical behaviour in order to effectively deal with such situations. If one can identify that self-deception was the cause, for example, then one can attempt to remedy the situation that gave rise to it so that any resultant harmful behaviour does not reoccur. Thus, I close Chapter 6 by looking at some of the suggested preventative measures for individual self-deception that I think could be applied to institutions. Additionally, I submit that one may be able to deal with the problem of institutional self-deception and an institution lying to itself through the introduction of different decision-making processes within institutions, in conjunction with a number of other administrative practices and other internal controls suggested by researchers such as Dan Ariely, a behavioural economist who has done a lot of work both on understanding and on curbing cheating, lying and deception in the corporate sphere which could provide some inspiration for improved oversight and transparency measures.

Chapter 1: What are Institutions, and Why Should We be Interested in Them?

> "In terms of power and influence you can forget about the church, forget politics. There is no more powerful institution in society than business." – Anita Roddick, *Business as Unusual* (2000)

Groups made up of people are highly variable – ranging, for example, from families and households to local chess clubs, book clubs, and neighbourhood watch groups, and from FIDE (the World Chess Federation), and the national police force, to NGOs, corporations, unions, churches, governments, and nations. We also think of Capetonians, vegetarians, or all left-handed people, everyone born the 27th of January, and women, as forming kinds of groups.

Given this wide array of groups, I will start this Chapter by looking more closely at what a number of theorists have said regarding such groups and the kinds of useful distinctions we can draw between them. In general, theorists tend to draw a very broad distinction between individuals that are grouped together on the one hand because they seem to have some feature in common with one another, or because they seem to occupy the same social caste etc., and groups on the other hand that are formed for some or other express purpose, or to achieve some end. I will then briefly discuss some of the strengths and some of the weaknesses of drawing such broad distinctions between types of groups, as well as the disadvantages that can arise for projects that seek to focus on very particular kinds of groups.

In spite of some of the disadvantages of doing so, I will be focusing on a particular kind of group – institutions – throughout my thesis. I put forward my own definition of what I take institutions to be in Section 2, and I will flesh this out in greater detail in Section 3. Institutions, as I define them, will be my focus because I think these kinds of groups have the internal structure necessary for mindedness and self-deception.[6] While I take institutions to fit into the broader category of groups which are formed for some or other purpose, I regard institutions as differing in a number of respects from other kinds of groups that could also fit into such a category.

I end in Section 4 with a brief discussion of why we should particularly be interested in institutions that wield an enormous amount of power and influence in our societies, although this will be further elaborated on in the forthcoming chapters.

[6] Although it may well turn out that other kinds of groups are also capable of mindedness and self-deception.

Section 1: On Grouping Groups

A number of theorists have attempted to isolate what they take to be the key features of all groups that are made up of people. For example, in her paper 'What are Groups?', Katherine Ritchie notes that all groups that are made up of people – what she calls social groups – seem to share the following six features: first, they can be made up of different members over time (i.e., person A can cease to be part of the group, and person B can become part of the group). Second, she suggests that social groups are capable of having different members in different possible worlds (i.e., while Group G is actually made up of members A, B and C, in another possible world Group G could have been made up of members D, E and F). Third she notes that social groups do not exist necessarily (i.e., they can exist at the moment, without having to exist at all possible moments). This extends to the thought that, fourth, they are capable of existing in one world without having to exist in all possible worlds (e.g., Google might exist in this world, but may have never been established in another possible world). Fifth she notes that social groups are located in space (i.e., because groups are made up of people who are located in space, groups are too – and thus it can make sense to say that the Rhodes Chess Club is currently in Port Elizabeth when its members are participating in a competition there). And finally, it is possible for social groups to coincide with one another and yet also be distinct from one another (e.g., it is possible that all members of the chess club, A, B, and C, also comprise all the members of a book club, and thus that there are two distinct groups made up of the exact same individuals) (Ritchie, 2013, p259).

Yet the groups listed at the beginning of this chapter illustrate that one might classify groups of people as such on the basis of geography, or preferences, or shared physical characteristics or other features, and alternatively, that such groups might be formed on the basis of certain individuals wanting to meet certain ends. Consequently, although they may share the above six features, groups can differ quite radically from each other in a number of ways. For example, one may think of all the left-handed people in the world as constituting a kind of group. Such a group, however, will be very different from, say, a jury, or a corporation. In the former, we have a set of individuals that can be grouped together because of certain physiological qualities that they have in common, while the latter examples refer to groups that have been be created to achieve certain ends – to decide the guilt or innocence of a party accused of breaking the law, or to provide a service or sell a product.

This very broad division between types of groups – roughly speaking, groups made up of individuals that share features versus groups created for a purpose – is one that has been

put forward by a number of theorists, although their specific terminology for distinguishing these groups differs. Margaret Gilbert, for example, draws what she takes to be an intuitive distinction between mere 'populations' (all left-handed people, or women) and 'social groups' or 'collectivities', saying: "the population of left-handed people, or the population of women, is not automatically a social group by virtue of having members with a common property […]. So much, I take it, is common ground" (Gilbert, 1989, p9). Christian List and Philip Pettit also draw a distinction between what they term 'mere collections' and 'groups' (List & Pettit, 2011, p31). For List and Pettit groups of left-handed people, for example, would be a mere collection, while the label 'group' is given to "purposive organizations" (p31). Deborah Tollefsen draws a similar distinction between what she calls 'aggregative groups' (a group classified as such on the basis of shared features, such as all left-handed people) and 'corporate groups' that have "a structure and a decision-making process" (Tollefsen, 2015, p3). Ritchie, who has written expressly on the nature of groups, has also drawn a distinction between these two types of groups which she labels Type 1, or "organized social groups" (2015, p9), and Type 2, or "feature-sharing groups" (2015, p13). As noted above, she thinks that all social groups share some common features, however she thinks that these two types of groups differ insofar as membership in feature-sharing groups, as the name suggests, relies on the individual members being regarded as sharing a particular feature with other members (being left-handed, say). In addition, Ritchie notes that members of feature-sharing groups do not usually (and certainly do not *necessarily*) have any shared purpose or intention with other members (2015, p8-9).[7] There is not usually any functional structure within such feature-sharing groups either, and she notes further that membership in feature groups is not obviously volitional.[8] In contrast, membership in organised groups like corporations tends to be voluntary (one chooses to try out for a sports team, say, or one applies for a job at a corporation), and organised groups are the kinds of groups that have a shared purpose or intention, and are structured in a particular way to achieve those ends, where to be a member of an organised group one needs to fulfil a certain role within the group (e.g., a goalkeeper in a sports team, or an accountant in a corporation) (2015, p9).

[7] Nevertheless, Ritchie notes that sometimes members of feature-sharing groups can develop a shared purpose. For example, Ritchie suggests that gender and racial groups should be regarded as feature rather than organised groups. Although Ritchie notes herself that many dispute that members of gender or racial groups do actually share any physical or psychological features (see for example Piper, 1992; Haslanger, 2000), they may still share socially constructed features with other members as a result of being labeled as *x* or seen as a member of the particular gender or racial group, and treated in a particular way as a result. The treatment of certain gender and racial groups can then lead to members of those groups sharing a purpose or intention with fellow members.

[8] Or at least it is often much more difficult to escape being categorised in a certain way because of perceived features – even though attempts may be successful on occasion.

There is enough overlap between the two categorisations of groups put forward by the authors above for the claim that this kind of broad division between groups is commonplace. Yet Brian Epstein (2017) is critical of what he regards as such oversimplified approaches to categorising and analysing groups. Epstein cautions that, given the incredible diversity of groups, it is simply naïve to suppose that we will be able to neatly divide such groups into two or even three categories or to suppose that, if we do, that we will have offered an account that adequately encapsulates all the groups we would actually take to be groups under the appropriate classification. For example, exactly what groups should be regarded as feature-sharing groups and what groups should be regarded as organised groups is something of a grey area – for example, is being a vegetarian something more akin to being born on the 27th of January where the same thing is true of all members (i.e., they don't eat meat), or more like being part of a chess team, where members observe certain rules or practices, perhaps with a similar end in mind?

For Epstein, the distinction which Ritchie and others make between feature-sharing groups (like all left-handed people) and purposive groups (like chess teams, corporations and governments) is already misleading because of the incredible variety of kinds of groups we encounter within each of those categories (Epstein, 2017, p2). In a more focused criticism of Ritchie's assertion that we can draw a neat distinction between two fundamental types of groups – what she has referred to as feature-sharing groups and organised groups – he argues that how we understand the term 'feature' can complicate things. Depending on whether one understands 'features' as just intrinsic properties, or as extrinsic properties too, one may end up either leaving groups out of the class 'feature groups' that one only thinks of as groups at all because of certain shared extrinsic features (e.g., Capetonians), or one may end up with many (if not all) organised groups *also* counting as feature groups (Epstein, 2017, p5). This would be by virtue of the fact that members of particular organised groups would share certain socially constructed features, such as 'being a board member of Google', or 'being a member of parliament' etc. (Epstein, 2017, p5). Epstein further notes that even if we could classify some or other group as a feature group, it can often be problematic to pick out just one feature and determine anything interesting about the group of people bearing that feature[9] – doing so simply leaves out so much information about other important details of the group (other features, or identity criteria to do with coming, or continuing, to exist, for example) (Epstein, 2017, p10).

[9] See Haslanger (2000) on the commonality problem and normativity problem.

Moreover, Epstein points out that many groups that seem to fit into Ritchie's category of organised groups, like committees, may often be made up of individuals that have divergent ends, such that they undermine one another rather than work together toward a joint commitment, or there may be organised groups like large corporations with many ends, or which have departments that no longer serve any purpose within the organisation but are still part of it. For example, one can imagine an organisation with a department that is in charge of looking after grounds and gardens. If the organisation moved to premises with no gardens, then this department would be defunct. It could well be the case that it continues to be a department within that organisation even though it no longer serves any purpose, overlooked by the administration for quite some time. Thus Ritchie's characterisation of organised groups as having a shared purpose may be too strong a requirement for many groups that we take to fall into this category (Epstein, 2017, p6).

Ritchie claims that we can distinguish between her two classes of social groups by looking at "the volition conditions members operate under" (Ritchie, 2015, p8), noting that in organised groups membership is volitional (Ritchie, 2015, p9). Yet Epstein offers the example of an organised group of slaves and masters working on a plantation, and points out that clearly the volitional nature of membership in organised groups can also fail to hold (Epstein, 2017, p6). The same could be said of the penal labour which prisoners perform today.

It may also be the case that an organised group can cease to be, even though it has replacement members that perform the same role, if those members did not become members of the group in an appropriate way – thus perhaps a national cricket team could cease to be a national cricket team if the batsman was not replaced by another professional teammate, but by a random member of the public who was picked out on the fly. Even though they would still form a cricket team, they would no longer form the *national* cricket team. Epstein explains this by suggesting that "many kinds of groups plausibly have essential origins" (Epstein, 2017, p8) – in other words, many groups need to be formed in a particular way, i.e., by being elected to the committee, say, or by members being selected by a board with the requisite authority to do so etc. Thus, for many organised groups members will not be fungible in the way that Ritchie suggests they could be.[10]

[10] Ritchie may be able to deal with some of these criticisms by suggesting that it may be the case that certain organised groups will require that their members have certain features – yet I take it that Epstein's point about a clear distinction between feature and organised groups running the risk of oversimplification still stands.

Epstein recommends an alternative approach we should take to an analysis of groups in order not to mischaracterise them and any conclusions we may draw from any overly simplistic characterisation. He suggests that we need to begin by being very specific about the kinds of groups we are looking at (e.g., 'national soccer teams', and not just 'soccer teams'), before fleshing out four complimentary profiles regarding whatever particular kind of group we are interested in. These four profiles are a group's 1) 'construction' profile, 2) it's 'extra essentials' profile, 3) it's 'anchor' profile, and 4) it's 'accident' profile (Epstein, 2017, p2-3). By doing this, Epstein thinks that we will be able to make accurate and useful observations about whatever kind of group we have chosen to examine.

Very briefly,[11] a groups 'construction' profile would capture various aspects of a group's construction such as the persistence conditions for the group, and what was necessary for the group to come into existence in the first place. A group's 'extra essentials' profile would include other aspects of a group that we take to be essential to it – for example, for certain kinds of groups it might be essential that it has certain rights or powers or responsibilities. A group's 'anchor' profile should provide information about what anchors its rights, powers, responsibilities etc. For example, some groups may have the abilities they do because they are legally recognised as having those abilities, while others may have the abilities that they do simply because of custom, because historically certain rights or roles have been given to these groups for various reasons (Epstein, 2017, p26; p38). And finally, a group's 'accident' profile will delineate accidental or non-essential properties, but properties which a group actually happens to have. This is something which could prove to be of interest when analysing a particular kind of groups – for example, that all members of a particular group are left-handed when they needn't be, or that all these groups are a certain size when they needn't be, could be of some interest to a number of concerns we may have when it comes to analysing groups (Epstein, 2017, p40).

As has been noted, it is important for Epstein that when applying this analysis one focuses on very specific kinds of groups – e.g., automobile manufacturers, or fast food corporations, and not just the broader category of 'corporations', as there are a range of different types of corporations that will obviously differ in some important respects from one another. The responsibility to create products that meet certain carbon dioxide emission levels is something that applies to automobile manufacturers, but not fast food corporations. Alternatively, if it turned out that every member of automobile manufacturers play tennis, or

[11] For a more detailed explanation of each of these profiles, see Epstein 2017.

that they are all white, or all male, this would seem to be an accidental feature of automobile manufacturers (as this is not something essential to automobile manufacturers). While that could be of interest when analysing that kind group, as it is an accidental feature it may not apply to the members of all other corporations. Nevertheless, Epstein also cautions against being too specific in one's analysis. He notes that we will likely fail to say anything of wide-reaching value if our focus is too small, as would be the case if we were to focus on analysing just one particular automobile manufacturer, like the Ford Motor Company, rather than on focus on the slightly broader group 'automobile manufacturers'. He acknowledges that we may well be interested in offering an analysis of a particular group, like the Ford Motor Company – how it came to be, what powers it may have, what structure it has – but notes that in order to do so comprehensively we would also need to bear in mind the kind of group of which the Ford Motor Company is a particular instance (Epstein, 2017, p9).

I am extremely sympathetic to some of the problems that Epstein highlights associated with breaking social groups into the very broad categorisations that many theorists do – even the very detailed account of two broad categories that Ritchie gives us. I also think that the extremely detailed analysis of very particular kinds of groups that Epstein calls for could be of immense value. It nevertheless is also the case that there are some descriptions we might want to use when talking about groups which could pick up a vast array of groups that differ quite radically from one another in terms of their 'construction' profile, 'extra essentials' profile, 'anchor' profile, and 'accident' profile. For example, one may wish to describe certain groups like chess teams, corporations and governments (while different in structure and ends and abilities, among numerous other features) as being minded, or as liars, or self-deceivers. As this is precisely my interest, I will be focusing on quite a general kind of group in my thesis – groups which I will call institutions. Below I outline exactly what I take institutions to be, and explain why this kind of group is too broad to fully flesh out the four profiles Epstein suggest one use when analysing groups.

Section 2: On Institutions

I take institutions to be organisations or societies founded for a particular purpose. In doing so I follow the definition of an institution found in the Oxford English Dictionary as "an establishment, organization, or association, instituted for the promotion of some object, esp. one of public or general utility, religious, charitable, educational, etc., e.g. a church, school, college, hospital, asylum, reformatory, mission, or the like" ("institution, n.", OED online). I

am therefore using the term 'institution' quite narrowly, as one may also find an additional definition for an institution in the OED as an "established law, custom, usage, practice, organization, or other element in the political or social life of a people; a regulative principle or convention subservient to the needs of an organized community or the general ends of civilization" ("institution, n.", OED online). In his paper, "What is an institution?", John Searle incorporates this additional understanding of institution into his account and talks of things like language, or marriage, as institutions (Searle, 2005). I'm not interested in as wide an understanding of institutions as this as my focus is on groups, and thus what could be seen as institutional entities – a kind of organised group established for some or other purpose, such as those listed in the first definition above: a church, school, college, hospital, asylum, reformatory, mission, or the like.

It will become apparent that, in narrowing my focus as I have, I am motivated in part by a question posed by Stephanie Collins. She asks:

> When we want to demand that our world gets better, to what sorts of group may we address that demand? Is it too ambitious to say that 'humanity' should solve global ills? Is it too reductive to say it's all down to individuals? (Collins, 2019, *introduction*)

There has been a great deal of focus by a number of philosophers on a range of questions surrounding groups that are created for a purpose (like chess teams, or corporations), because when questions like that of Collins are posed, these are the kinds of groups that we often tend to think most capable of acting efficaciously in our societies, and indeed have come to regard as responsible in a variety of ways for a number of social ills (and not the sorts of collections of individuals one could group together because of some perceived commonality – such as being left-handed, or being born on the 27th of January).[12]

Institutions are the sorts of entities that are explicitly established, or 'instituted', as entities. This means that they will often be legally recognised as an entity (e.g., a business that is not a sole proprietorship will have to be registered as such with the state in countries like South Africa) [13] – although it will not always be the case that they are so recognised. In fact, some institutions like opposition political parties are often recognised as being illegal. Recognition of an institution as an entity is nevertheless significant – in recent years an

[12] Gender and race could be regarded as exceptions here.
[13] Indeed, in some countries many institutions are even recognised as persons - see for example the official South African government website which states that "the law sees a company in the same light as a natural person" (South Africa Government, 2021).

increasing number of institutions across a wide range of sectors are being held legally responsible for perceived wrongdoing.

In addition to the above, institutions are groups that are typically hierarchical, and have a division of labour in order to effectively meet their established purpose. Thus members of institutions will occupy certain positions arranged hierarchically according to power, status, and their functional role within the institution. For example, a board of directors will usually sit at the top of the hierarchy in a corporation like Ford Motor Company. The executive officers of the company will occupy the next highest tier of such a company, led by the CEO. Other levels of institutional hierarchy would include managers of different branches within the company which focus on different tasks that serve the overall aims of the institution, like that of marketing, safety, human resources etc. In most institutions, the higher one's hierarchical position the greater the influence one will wield when it comes to policy and decision-making etc.

Finally, institutions are the kind of groups that are governed by certain rules and norms which will be derived from both the structure of the particular institution, and whatever purpose the institution was founded to pursue – e.g., schools will have codes of conduct that staff and students are expected to adhere to, but they will also be expected to perform their function of educating children, while car rental agencies will also have internal rules that clarify the roles and responsibilities of individual employees, and penalties for misconduct, while at the same time they will need to fulfil their function of renting cars.

Institutions will thus differ from what Gilbert calls 'plural subjects' – something which at first sounds as if it could count as an institution. Indeed, Gilbert explains that "people form a plural subject […] when they are jointly committed to doing something as a body, in a broad sense of 'do'" (Gilbert, 2000, p2), a definition which seems to match up quite nicely with the definition of an institution as a group or organisation formed for a particular purpose. However, for Gilbert two people walking together are a paradigmatic plural subject (Gilbert, 1990). Yet because two people walking together are unlikely to have been explicitly or formally established as an entity (at least not without some other stronger connection between the two, like a business contract), I would not regard them as an institution. Similarly, a group of friends who are playing a game together would not constitute an institution, since they will be unlikely to have explicitly or formally established themselves as a game-playing entity, and all else being equal, others would be unlikely to recognise them as such. Given that institutions are also hierarchical with a division of labour, that such a group of friends has a shared purpose, and even a joint agreement to play the

game together, is not sufficient for an institution. A hiking club, on the other hand, even if it were comprised of just two people who walked together, could be a plausible candidate for being regarded as an institution, as could a games club made up of friends.

With the above in mind, I also regard 'institutions' as slightly different to Ritchie's organised groups. She sees organised groups as having "structural-functional organization" (Ritchie, 2015, p9), which would seem to accord with my view of institutions as entities tending towards hierarchy and a division of labour. She notes too that when individuals want to form an organised group, like a team, or a committee, they must have a shared intention to do so, and as such, it would seem that my view that institutions are explicitly established could also be met. I am not sure, however, that it would be necessary for Ritchie's organised groups that outsiders recognised them as an entity – yet I regard this as an important feature of institutions. As noted above, institutions usually tend to be recognised (or at least recognisable) as entities by the state. Additionally, going back to Epstien's critique regarding the volitional (or not) membership of organised groups, I would also regard prisons as forming institutions – it would seem that such structured groups need to be explicitly and formally set up (they don't just emerge spontaneously or organically from the interactions between people), they involve hierarchy and a division of labour, and they are governed by rules and norms. Yet given that membership within these kinds of institutions is not always volitional, this is another feature of Ritchie's organised groups that I am not convinced institutions will always share. With regard to prisons at least one can presume that, although they participate in certain rehabilitation programmes, and perhaps organised penal labour, prisoners would lack shared intentions to establish or be part of the institution of which they have become members. Consequently, institutions should be seen as a different category to Ritchie's organised groups.

Epstein would argue that my category of institutions is too broad. There are a wide variety of institutions that are established for a range of different purposes, such as social, religious, educational, or professional purposes, and thus 'institutions' is a broad enough category to incorporate both private security companies and the national police force – and thus not particular enough for Epstein – as well as a range of other institutional groups that differ even more from these, such as automobile manufacturers, universities or governments. As a result, while I will attempt to do so in Section 3 below, I will not be able to flesh out Epstein's four profiles in detail when it comes to the group 'institutions' because of the immense range of different sorts of institutions one can have, and so their 'construction', 'extra essential', 'anchor' and 'accident' profiles would inevitably differ quite radically from

one another (although such an analysis could be carried out for one particular institution, or particular kinds of institutions such as automobile manufacturers).

I will nevertheless be focusing on the broader category of institutions in this thesis because I think that these kinds of groups have the potential to be minded, and thus self-deceived, regardless of whether a particular institution is formed via members consenting to be part of it, while members of another institution are forced or voted in. These may well be relevant factors for a number of other interests we may have,[14] but these factors will not be relevant to my own broader concerns here. Moreover, I do not want to limit the scope of potential applicability.

I am interested in groups that have the potential for self-deception – *as* groups. Institutions are a promising candidate in this respect. As noted, institutions are founded for a particular purpose. In working to achieve its purpose, when an institution has division of labor, hierarchy, rule-following, etc., the conditions are created such that it becomes crucial to see it as minded, and in some cases, self-deceived.[15]

While it may be possible that a number of other kinds of groups could also be susceptible to self-deception, such groups will not be my focus at this time. Moreover, it need not be a problem that certain kinds of groups which have no aims (e.g., groups made up of all left-handed people) will not be capable of self-deception.[16] We need not suppose when examining some or other group level phenomena that *all* groups will necessarily give rise to them.

[14] It is possible that a very detailed understanding of a subset of institutions (like automobile manufacturers), or of a particular institution (like the Ford Motor Company) would be of interest once one has identified an institution that is guilty of wrongdoing. Such a detailed analysis could lead to a better understanding of what led to the wrongdoing, or how to implement positive changes in certain industries.

[15] A number of theorists have suggested that a much wider range of groups could be seen to act as groups and even have a 'group mind' – such as mobs, for example (Le Bon, 1985).

[16] It may well turn out that a broader range of groups that I am considering here will be capable of self-deception (e.g., nations) – however I do not have the space to explore this possibility here. Ian Deweese-Boyd (*forthcoming*), for example, explicitly suggests that a non-institutional group like 'consumers' could be capable of what he terms 'collective self-deception'. Yet what he seems to have in mind is that this collection of individuals all share similar self-deceptive beliefs about their consumer habits, as opposed to a stronger view about the group of consumers itself manifesting self-deception – a group which, lacking a cohesive purpose, strikes me as an implausible candidate for self-deception. Conversely, it may well turn out that a number of groups one might not ordinarily think of as 'institutions' end up meeting my criteria for so counting – such as families. This may not be too odd when one considers more extreme examples of families like mafias, or a royal family, say, and considers the parallels between these and other more run-of-the-mill families.

Section 3: A Look at Institutions in Greater Detail

To illustrate both of the points made above, namely a) that institutions form too broad a category for a single analysis making use of Epstein's four profiles listed above, but also that b) institutions as a category do share features that are important for my own concerns regarding a potential for mindedness and self-deception, it may be worthwhile to see what can be said of the fairly broad category of institutions in terms of Epstein's suggested profiles.

In terms of their construction profile (i.e., what is necessary for the group to come into existence, the persistence conditions for the group etc.), institutions are the kinds of groups that are intentionally and explicitly set up to meet some or other end – a small business may be set up to buy and sell used cars, an automobile manufacturer may be set up to manufacture and sell cars, while a public library might be established to provide the public with free access to reading materials, and universities may be established to provide tertiary education to members of the public (whether or not an institution succeeds in meeting their end is not of concern here). Oftentimes institutions will come into existence after a number of legal documents have been submitted in order to register the company or public institution and these documents been approved by the relevant authority, although the details involved can differ to some degree. Institutions also seem to be the kinds of groups that "exist in perpetuity or until they are expressly disbanded" (Epstein, 2017, p20) – i.e., the loss of employees at a company, or even the board of directors or CEO of a company, does not necessarily entail the end of the company. It is conceivable that an institutional entity could even lose all of its members and continue to exist.[17]

In terms of what it is that makes an institution the same institution over time is a bit more complex. Given that institutions are officially or legally recognised entities, institutions can potentially survive in spite of a complete changeover of membership – as long as any new or replacement members meet the requirements for being part of that institution, they form the same institution. If the second-hand car dealership gets sold and the new owner brings in a completely new contingent of staff, it is still the same business. It also does not seem that the new members of an institution need to perform the same roles as old members

[17] This is likely a very rare occurrence, but technically possible – for example, if everyone who worked for a registered company died in a plane crash, the company would still exist on paper, and would need to be dissolved. This doesn't seem to be something that would be true of a group like a book club - if all the members of a book club were to die in a plane crash, unless the book club is also an officially registered club, the book club ceases to exist without them.

because institutions can grow (a business can add another sector, or a university can introduce a new department). They can even potentially change in what might seem to be a fundamental way (for example, the distillery based in London that specialised in making craft gin stopped making gin and started producing hand sanitizer instead in March 2020 [Miller, 2020] – just one example of an institution that fundamentally changed its purpose without ceasing to be the same institution).

In terms of the extra essentials profile (which covers other aspects of a group that we take to be essential to it), institutions are established for a purpose, hierarchical in nature, and make use of a division of labour. As institutions are usually legally recognised entities, they can also have a variety of legal powers and abilities – they are able to enter into legal agreements, they can often sue and be sued by others etc. Yet it is also the case that the particular powers and abilities an institution will have will vary from other institutions, depending on their purpose. For example, a university (or a department within a university) may have the ability to fail a student who does not meet certain academic requirements, and a government may have the power to call a national election. Yet these would not be abilities that other institutions like hospitals or car manufacturers would have. As such, this profile is one that is difficult to flesh out in any detail for institutions. Yet in more general terms, we can regard institutions as the kinds of entities that are able to undertake a range of tasks (like hosting a speaker, or advertising, or refusing entry to members of the public, or making promises, or calling elections etc.).

The agreed purpose of an institution, and the fact that many institutions are officially if not legally recognised, could be seen as what generally anchors an institution and gives rise to the kinds of abilities that it has. Furthermore, the structure of an institution – that they tend to be hierarchical and have a division of labour – will also ground its ability to carry out its tasks.

The accidental properties of institutions will also vary from institution to institution – and not just broadly between kinds of institutions, but also between institutions of the same kind. For example, two businesses that provide the same services may have a different number of employees, or may have put in place different kinds of decision-making processes. Some businesses may happen to only employ women, or happen to be made up entirely of people who play golf over the weekend, or who happen to have been born in a particular country. This profile then is particularly problematic to fill out for the large category of institutions – there are just too many different kinds of institutions to discern a noteworthy

pattern. Thus, it is unlikely that one will be able to make any claims about the larger category of institutions here.

With this in mind, we can see that institutions cannot be entirely satisfactorily analysed in terms of Epstein's four profiles. Yet attempting to offer a detailed analysis of this broad category does highlight features that institutions, no matter how different their ends, do all have in common, namely that they are purposive, recognisable as entities, and have a hierarchical division of labour. As such the kinds of problems that one may encounter with regard to the behaviour of different institutions like car manufacturers, universities or governments could have a very similar source, and perhaps a very similar solution. Moreover, making use of Epstein's four profiles could still yield very valuable insights when it comes to analysing a very particular set of institutions (e.g., car manufacturers), or indeed a single institution (e.g., Ford Motor Company).

Section 4: Why be Interested in Institutions?

That we treat institutions as if they are agents is an observation made by a number of authors writing in this field. As Collins notes, "to find evidence of this, you just need to open a newspaper" (Collins, 2019, p1). Following her suggestion, we can look for example at some headlines from a number of different news outlets:

> Spy poisoning: Russia expels 60 US diplomats in tit-for-tat measure (*BBC News*, 2018)
> DA left embarrassed as Patricia De Lille survives motion of no confidence (*The South African*, 2018)
> Tiger Brands Refuses to Take Blame for Listeriosis Deaths (*Eyewitness News*, 2018)
> Google Thinks Self-Driving Cars Will Be Ideal For Senior Citizens (*TechTimes*, 2016)

Here we have institutions responding to insults, feeling embarrassed, refusing blame and thinking things. These headlines are suggestive of the fact that we see companies such as Google, or political parties like the DA as minded, and as capable of intentional action in the world. As noted, Collins is not the only one to make such an observation. Tollefsen and List and Pettit have the following to say:

> In everyday contexts we often talk as if groups were agents with attitudes such as belief and intention […] economists and political scientists talk of group rationality; sociologists, historians and anthropologists often appeal to the concept of a group's memory; […] organizational theorists study the ways that firms and organizations learn and remember (Tollefsen, 2015, p1-2)

> Common sense and the social sciences represent many collections of human beings as if they were unitary agents, capable of performing like individuals. We speak in ordinary life of what Greenpeace or Amnesty International intends, what the Catholic Church holds sacred […]. (List & Pettit, 2011, p1)

What's more, institutions in particular are often capable of wielding an enormous amount of power – just think of influence entities like the government, the church, or large corporations have over our lives (and over each other). Governments are making and enforcing laws which have an impact on our daily lives, for example, and corporations can lobby governments to influence their decisions in their own favour. Even small institutions are likely to have greater reach than the average individual in terms of being able to achieve their ends – a famous speaker is more likely to attend an event organised by a local school than one organised by an individual.[18] Institutions, then, are some of the most powerful social groups on the planet – and so they have the potential to cause an enormous amount of good, but also an enormous amount of harm.

From at least one of the headlines above it should be clear that institutions cause harm. If Tiger Brands had been less negligent, perhaps, it is possible that a number of deaths could have been avoided. In forthcoming chapters I will argue that the internal structure of institutions creates plausible conditions for mindedness. If institutions are minded, then we need to understand how exactly they are causing harm. It is possible that the harm they cause is deliberate and intentional. Sometimes it may be the result of negligence. However, it is also possible that the harm which institutions cause can come about because of self-deception. Since it is widely acknowledged that self-deception is at least potentially dangerous when individuals succumb to it, given that institutions tend to be able to wield far more widespread power and influence than individuals, the potential danger or damage of self-deception may be increased if an institution is self-deceived. I will argue that the potential harms of institutional lying – intentional wrongdoing – and those of institutional self-deception are comparable.

Moreover, if institutions do cause harm, attributing responsibility for that harm becomes relevant. A number of the features highlighted in the above analysis will be of interest to those concerned with holding institutions responsible for harmful behavior. Institutions are hierarchical entities established specifically in order to meet certain ends,

[18] This will not necessarily always be the case. An extremely famous individual may have more clout than a school that is not well known, or one with a bad reputation.

where a division of labour contributes to the effective pursuit of those ends. Yet institutions are also capable of persisting over time in such a way that they become separable to some degree from their individual members (i.e., the institution persists despite changes in membership). We have seen that institutions are also the kinds of entities that can survive (radical) changes in membership, which may well be of interest when it comes to addressing the actions of institutions in our lives. Nevertheless, we do see them as discrete entities, responsible for various harms. Before we can say anything meaningful with regard such concerns, however, the question of how exactly we ought to make sense of group mind talk, and how we ought to understand the relationship between institutions and their individual members, needs to be addressed. I do so in the following chapter.

Chapter 2: Making Sense of 'Group Mind' Talk When it Comes to Institutions

> "Google Thinks Self-Driving Cars Will Be Ideal For Senior Citizens" -
> Mark Lelinwalla, *TechTimes* (2016)

At the end of Chapter 1 I drew attention to the fact that we do talk of groups such as institutions as if they were minded, intentional agents and, moreover, that such talk is ubiquitous. However, the question of how seriously we ought to be taking our tendency to see institutions as having mental states, or to see them as agents, needs to be considered. I begin in Section 1 by addressing some preliminary concerns about the suggestion that certain kinds of groups are minded – concerns of both a) an ontological nature, where group-level mental states are regarded as reducing to the mental states of individuals, and b) an explanatory or methodological nature, which involve claims about what the best explanations are that we have on offer regarding the mentality or agency of groups like institutions. I will then look at one of the most compelling reasons we have to favour the suggestion that institutions are minded, or have mental states of their own in Section 2. In doing so I follow Tim Crane's suggestion that we ought to regard the mind as a range of mental capacities, i.e., that to have a mind is "to think, to act, to feel, and to have emotions and to be conscious" (Crane, qtd. In 2010, p135), as opposed to seeing the mind as a particular kind of *thing*. For Daniel Dennett, to have a mind is to be creature towards which we can successfully adopt the intentional stance, which involves taking that creature to have certain mental capacities in order to be able to make sense of what it does. When writing headlines about the thoughts and reactions to certain events of corporations and political parties, it seems that this is precisely what we do as a way of making sense of these institutions. With this practice of ours in mind, I will look at the argument that we can very successfully adopt an interpretive or intentional stance towards groups like institutions in Section 3.

I conclude that the success of applying a functionalist view of the mind like Crane's to institutions, as well as the success of adopting the intentional stance toward institutions, are indicative of mindedness – although I will not be arguing here that institutions are the kinds of things that have all the mental capacities we might be inclined to attribute to a mind (like consciousness for example). Yet I take it that these successes will not be enough to convince the sceptic that institutions are in fact minded. Such arguments only allow for the possibility of institutional minds. Nevertheless, such views will pave the way for taking Gilbert and List and Pettit's explicit arguments for the existence of group mental states seriously. In Chapter 3

I will outline their arguments in more detail, and go on to suggest that not only can we see institutions as genuinely minded, we ought to see them as moral agents too.

Section 1: Scepticism about Group Minds

There are two concerns about the reality of group minds or group-level mental states that I will discuss – namely, an ontological concern and an explanatory or methodological one. When it comes to talk of group mental states, one criticism has been that even though we may naturally talk of institutions as thinking, feeling, and acting in the world, we should recognise that such tendencies are simply metaphorical shorthand for attributing these states to the individuals that make up those institutions. There are a number of different ways of cashing out this view. For example, on the more extreme end of the spectrum one could mean that while it may seem obvious that groups like institutions exist, any reference to such groups is merely convenient metaphorical shorthand for referring to numbers of individuals. At the end of the day, all talk of groups like institutions as thinking, feeling and doing things is a kind of fiction – all we're really talking about is individual human beings. This kind of view has been labeled eliminativism, and is contrasted with realism about groups and group minds.[19]

Katherine Ritchie points out that one can be a realist about some kinds of groups and an eliminativist about others – for example, one might be happy to concede that groups like sports teams exist, but be an eliminativist about other kinds of groups, like racial groups (Ritchie, 2015, p2). Or, as Ritchie points out, "one might be a Group Eliminativist about groups that have agency or mental states, while being a Group Realist about non-agential unconscious groups" (Ritchie, 2015, p2). In other words, one may not think that groups can have mental states or agency, but think that groups nevertheless exist.[20] I take it that the existence of institutions is not disputed.

One can also think of talk of certain group's having mental states and performing actions as merely metaphorical without subscribing to such a strong ontological claim about the existence (or lack thereof) of groups *per se*. In this respect, Anthony Quinton famously asserts:

> We do, of course, speak freely of the mental properties and acts of a group in the way we do of individual people. Groups are said to have beliefs, emotions, and attitudes and to take decisions and make promises. But these ways of speaking are plainly metaphorical. To ascribe mental predicates to a group is always an indirect way of ascribing such predicates to

[19] See List and Pettit (2011) and Ritchie (2015) for a more detailed discussion of eliminativism.
[20] See Robert Rupert (2014) who defends such a view regarding the existence of groups with mental states.

> its members . . . To say that the industrial working class is determined to resist anti-trade union laws is to say that all or most industrial workers are so minded. (Quinton, 1975, p17)

Quinton appears to be an eliminativist about group minds in saying that attributing mental states to groups is "always" an indirect way of attributing them to the members of that group. Consequently, it would seem that he does not think that mental states are the kinds of things that groups could ever have. This approach can certainly be appropriate when it comes to the ascriptions of mental states to certain groups. With regard to Ritchie's feature groups, when talking about what women, for example, want, or believe, it may be the case that this is genuinely better understood as just shorthand for talking about what all, or most individual women want, or believe. Yet there could be other scenarios in which this may not be the case, as, I will suggest, when talking about what institutions want or believe.

Yet many theorists who do subscribe to views of (certain) groups as thinking agents in their own right are happy to accept an ontological reductionism of groups insofar as it is regarded as the claim that such groups are made up of individuals (rather than that groups don't exist at all). For instance, when talking about groups of people that are jointly committed to some or other end, or 'plural subjects', Gilbert explains that:

> plural subject theory […] goes beyond individualism both with respect to the concepts it uses and in its understanding of what there is. It is worth emphasizing that this does not mean that plural subject theory goes beyond the plane of humanity in characterizing human sociality. It does not invoke any ontologically suspect kind of "social spirit" or "group mind." Human beings create joint commitments together and thereby constitute plural subjects. (Gilbert, 2000, p3)

Gilbert herself is reluctant to use the term 'group mind', yet talks about the possibility of 'group beliefs' (1987; 2000) and other collective states we would typically think of as mental or indicative of mindedness – and does so while acknowledging that such states supervene on individuals, albeit in a certain kind of relationship with one another. For example, Gilbert has later stated that "collective agents, as I understand them, act through their members" (Gilbert & Pilchman, 2014, p89), emphasising again that while there may be group-level states, they will supervene on some combination of individual contributions. List and Pettit make a similar point when they say of their view about group agents, "we put aside the mystery associated with the emergentist tradition in building our account of group agency on the thesis that this phenomenon supervenes on the contributions of individuals" (2011, p74-5). Deborah Tollefsen, too, notes quite simply that "groups are composed of individuals, and they act via the actions of individuals" (2015, p4), indicating an acceptance that groups and

their actions etc. are in a sense dependent on the individuals that make them up. Such theorists take issue when these facts are then taken to imply that there can therefore be no group-level properties that individuals within the group do not individually possess.

Another concern that one might have with regard to positing group mental states comes from an appeal to methodological or explanatory reductionism. According to Marie Kaiser, methodological reductionism is a normative claim about how research in a specific field ought to be carried out, and the argument is that "research should be pursued in a reductionist fashion because this is the most successful way to pursue it" (Kaiser, 2015, p72). In slightly more extreme phrasing, Kaiser notes that for the methodological reductionist, we should "always seek explanations by investigating the under lying processes at lower levels of complexity, and ultimately at the level of atoms and molecules" (Ayala, qtd in Kaiser, 2015, p71-2). While one might not wish to reduce explanations of group level phenomena which make reference to group level mental states all the way down to explanations at the level of atoms and molecules, one might nevertheless argue that when it comes to understanding the social world, and understanding how and why institutions behave in certain ways, one will be most successful at doing so if one adopts a reductionist strategy and attempts to work out what is going on at an individual level within these groups. This position is referred to as 'methodological individualism', and Tollefsen notes that such a view entails the belief that "groups and their actions can be explained solely in terms of the psychological states and processes of group members and the relations between them" (Tollefsen, 2015, p4).

In 'Group Minds and Explanatory Simplicity', Mark Sprevak and David Statham discuss an argument that appeals to simplicity against positing explanations that make reference to properties, like mental states, at the group level. One of the main implications of the argument is that "for any explanation of intelligent behaviour that appeals to group minds, there exists an alternative explanation couched solely in terms of the minds of individuals" (Sprevak & Statham, 2015, p1). Positing group level mental states, or group minds, violates some variant of Occam's razor. Although in favour of certain attributions of group level mental states, List and Pettit argue for this too – when offering an explanation, one should adhere to the principle that one ought not multiply entities beyond necessity, thus if one is able to offer a perfectly adequate explanation in terms of only individuals and their mental states, one should not posit something additional – namely group mental states or minds (List & Pettit, 2011, p2). Thus, there is a normative dimension to the position of methodological or explanatory reductionism. An explanation that does not require references to group-level

metal states is surely a simpler explanation than one that does – and "simpler explanations are, everything else being equal, better explanations" (Sprevak & Statham, 2015, p1). Therefore, the argument goes, "individual-level explanations are, everything else being equal, preferable to their more complex group-level alternatives" (Sprevak & Statham, 2015, p1).

When thinking about what kinds of explanations we *ought* to be offering, again one may think that although we can technically offer explanations that make reference to groups or group-level properties, we ought to be offering reductive explanations of social phenomena in terms of just individuals and their individual properties because that is all that there really is – i.e., one would hold that group level explanations, or references to group level properties, are metaphorical and so, while perhaps useful in just the way that metaphorical explanations are often useful they nevertheless fail to be, strictly speaking, true. Yet it could also be that one thinks that explanations in terms of individuals are superior, and therefore the kinds of explanations we ought to strive for, because they are either simpler, more straightforward, or because they have some other admirable trait that explanations at the group level fail to have[21] – e.g., one may think individualistic explanations are more complete than explanations at the group level because one thinks that in order to explain what is really going on at the group level one will need to revert to explanations of individuals, their properties, and their inter-relations.

It is worth noting that there may be some tension between ontological reductionism and methodological non-reductionism – one can be a reductionist about things, but not about properties. In other words, one can subscribe to ontological individualism insofar as one distances oneself, as List and Pettit do, from "any appeal to psychologically mysterious social forces" (List & Pettit, 2011, p3) – without necessarily thinking that the best explanations of the behavior of groups like institutions will be offered solely at the individual level. This requires accepting the claim that "there are irreducible group-level properties and processes which need to play a role in the explanation of group phenomena" (Tollefsen, 2015, p4). I am interested in precisely such a position – that institutions, although made up of individuals, and dependent on individuals and their attributes and inter-relations, are capable of having irreducible beliefs, and, as I will argue, self-deceptive beliefs, of their own. Philosophers like List and Pettit, and Gilbert have put forward a number of arguments for irreducible group-level mental states, while others like Tollefsen provide us with arguments for taking these

[21] Although it is inevitably more difficult than it may seem to make these sorts of claims, because the notion of explanation is a multifaceted one, and as Samir Okasha notes, "doubts have been raised about whether […] the concept of 'best explanation' can be made precise" (p692).

ascriptions of group level mental states seriously. Before looking at the arguments put forward by Tollefsen in this Chapter, and List and Pettit, and Gilbert in the next, I will look briefly at how the suggestion that institutions could be minded can get off the ground.

Section 2: On Mind and Group Mind

Tim Crane and Jaegwon Kim are both advocates of seeing the mind as a set of mental capacities rather than any particular kind of substance. Indeed, Kim explains that "having a mind can be construed simply as having a certain group of properties, features, and capacities that are possessed by humans and some higher animals but absent in things like rocks and trees" (Kim, 2010, p3). Given the problems that have arisen for substance dualism as well as physicalism,[22] this would appear to be a promising way of understanding what it is to have a mind.

Nevertheless, one particular difficulty one may have with Crane and Kim's suggestion is trying to identify what exactly is meant by the term 'mental' capacity. Crane offers some examples: "the capacity to think, to act, to feel, and to have emotions and to be conscious: all these mental capacities are the things that I would classify together as the mind" (Crane, 2010, p135). Kim talks about the possibility of isolating some or other feature in that list that could prove to be the unifying feature of all mental capacities. Ultimately, however, Kim suggests that this is a notoriously difficult thing to do, and that to date none of the potential unifying candidates that have been proposed are successful in actually unifying all we take to be mental: "none seems to be adequate for all the diverse kinds of events and states, functions and capacities, that we normally classify as 'mental' or 'psychological'" (Kim, 2010, p8). My focus will be on an institution's ability to act and to possess certain mental states like desires, intentions, beliefs and self-deception – states which I take it belong in the realm of the mental, as evidenced by Crane's list. I will argue that institutions can be minded in the sense that they are capable of acting and possessing certain mental states.

Functionalism as a theory of mind offers a different approach to substance dualism (where the mind is regarded as a non-physical substance) or type physicalism (where the mind tends to be identified with a particular physical substance, like the brain), insofar as it suggests that mental states ought to be defined in terms of their functional role (where this

[22] As I do not have the space to flesh out these problems in detail here, see Kim 2000 and Ravenscroft 2005, among others.

includes references to inputs, outputs, and relation to other internal states), rather than in terms of the substance that plays those roles. Part of the appeal of functionalism is that it remains neutral regarding the make-up of the mental – as Ian Ravenscroft notes, although most contemporary philosophers of mind are physicalists, technically "it is in fact possible to be a functionalist and a substance dualist" (Ravenscroft, 2005, p53). One reason to favour a functionalist view of the mind is that functionalism can very easily accommodate the suggestion that mental states may be multiply realisable. For example, if one is not a functionalist but concedes that both a human being and an alien with a radically different physical make-up can both be experiencing the exact same mental state, one is presented with a problem. The belief that it is raining, for example, cannot obviously be reduced to some particular physical state, as the human being and the alien, who both believe that it is raining, will be in different physical states (due in this case to their radically different physical make-up). Here it may be useful to make use of a well-worn software/hardware analogy: functionalism opens up the possibility that, metaphorically speaking, the exact same 'software' (e.g., the mind/mental states) could be run on many different types of 'hardware' (e.g., a human brain, an animal brain, an alien brain, AI, or, conceivably, an institution etc.).[23]

Some physicalists have attempted to counter the problems posed by multiple realisability by suggesting that while some or other mental state like the belief that it is raining might be differently instantiated by a human being and by an alien, all mental states will nevertheless reduce to some or other token physical state. While this may well be the case, one potential weakness with this token reductionism is that it may become difficult to say in virtue of what X and Y should both count as instances of the same mental state, such as 'the belief that it is raining', given that the physical bases for the mental state are so different. In other words, one may well wonder what it is that makes, say, h-fibres firing in a human and a-fibres firing in an alien both instances of a belief that it is raining. Once we identify particular mental states in terms of the functional role that they play in the system of which they are a part, one can say that both X and Y are instances of the same mental state because they both perform the same role within their respective systems.

If to have a mind is to have a certain set of capacities, and if we can understand these capacities as the ability to fulfil some or other function, then it is possible that these capacities can be multiply realised. Moreover, if functionalism puts no constraints on the type

[23] The hardware that instantiates mental states could also, of course, conceivably be the soul.

of base that can realise a mental state, then it opens up the possibility that institutions could be the kind of base that can realise a mental state.[24]

Nevertheless, there are a number of problems that arise for a functionalist account of the mental. Nagel draws attention to the failure of functionalism to accurately capture or account for conscious experience, saying that "there is something it is like to be that organism […] It is not analyzable in terms of any explanatory system of functional states, or intentional states, since these could be ascribed to robots or automata that behaved like people though they experienced nothing" (Nagel, 1974, p436). Thus, although functionalism is meant to be able to characterise mental states, like pain for example, in terms of their function – which includes reference to certain inputs (in this case perhaps bodily damage), and certain outputs (attempting to withdraw one's body from the source of pain, or an inclination to yell in pain, etc.), and even reference to other internal, mental, states (like the belief that 'I am in pain', etc.) – it is not obvious that functionalism will be able to account for qualia. It might seem that institutions could feel pain – a very recent headline, for example, reads: "Small Businesses Continue To Suffer Pandemic Pain. Is Bankruptcy The Answer?" (Frankel, 2020). Thus a business might respond to certain inputs that are harmful to their ability to operate like restrictions on their opening hours or services with certain outputs like closing down branches, or implementing layoffs. Yet it is also possible that an institution may fail to actually *feel* any pain, even though they can arguably behave as if they did. If the institution did not feel any pain, however, then one would question whether or not it was in pain – and so had a certain mental state – at all.[25] Daniel Dennett has written about our incredible susceptibility to being 'charmed by the illusion of consciousness', even in things we know for a fact aren't conscious. He cites his own involvement in the programming of a robot, 'Cog', at MIT, and notes that even though no programming of any degree of complexity that could plausibly give rise to consciousness had taken place, because the robot was endowed with eyes and hands etc., all those who were working on Cog couldn't help but feel that they were "in the presence of another conscious observer when Cog's eyes still quite blindly and stupidly follow[ed] one's hand gestures" (Dennett, 1995, p694).

[24] Here it is important to note that functionalism just opens up the possibility that institutions could instantiate minds – functionalism itself is not a direct argument for institutional minds, but it allows for that possibility to be taken seriously.

[25] Arguments to a similar effect have been made against functionalism – see for example Ned Block's (1990) discussion of functionalism's inability to account for different phenomenal experience in an inverted spectra scenario, and thus its failure to successfully capture an important aspect of certain mental states, namely their qualitative feel.

An additional critique of functionalism is that it fails to genuinely capture what could be thought of as understanding or language comprehension, witnessed by the fact that a symbol-manipulating machine could plausibly pass a Turing Test. Given certain inputs, such a machine could deliver the 'correct' outputs – yet as John Searle objects, sophisticated robots or machines of this sort would only appear to display any understanding of what was being asked of it. Searle famously makes use of the 'Chinese Room' thought experiment to illustrate as much, whereby someone who has no understanding of Chinese could nevertheless appear to understand the language simply by matching up symbols and selecting responses in a rule book (Searle, 1980, p417-418). This kind of criticism was previously expanded upon in the 'China Brain' thought experiment proposed by Ned Block, in which each of the billion citizens of China manipulate symbols for an hour in a highly organised way, and China as a whole thus behaves appropriately in exactly the same way as the sophisticated robot was able to – answering questions, expressing beliefs, and being "functionally equivalent to you for a short time" (Block, 1978, p279). The intuition here is meant to be not only that 'China Brain' would fail to display any genuine understanding but, moreover, that we should doubt such a "homunculi-headed system […] has any mental states at all" (Block, 1978, p281). I would like to briefly address first the criticism that functionalism fails to show that certain entities that behave in a functionally appropriate manner have any genuine understanding or language comprehension before turning again to the concern that functionalism fails to account for qualia, something that is taken to be an essential feature of a number of mental states.

The Turing Test is originally set out in such a way that the interrogator who is meant to determine whether or not they are interacting with a robot or a person, and thus whether or not a machine can think, is set apart from the participant who answers questions posed by the interrogator. These questions are posed via some method of communication that requires no direct contact between the two – print-outs, an intermediary, or internet chat, say – presumably because to see the robot in action would give the game away (Turing, 1950, p433-434). With the Chinese Room thought experiment in mind, in scenarios such as this we may well be justified in being reluctant to ascribe any genuine understanding to a robot that could successfully pass the test. It could just be mindlessly manipulating symbols. However, we can imagine a number of situations in which we are not simply interacting with the robot via a keyboard in a different room. Rather, we are actually also watching the robot interact in the world – something which we also do with institutions. It seems that if the robot can pick out the yellow lemons from an array of various fruit and vegetables and give them to me

when asked to do so, then perhaps we ought to be prepared to say that it understands what is meant by 'yellow' and 'lemon', at least to some extent. Similarly, institutions can stay abreast of what's fashionable or topical, and respond accordingly. A chain of book stores, taking stock of the doom and gloom associated with current events can move its apocalyptic fiction to the current affairs section of its stores – and post a witty comment about doing so on a platform like Facebook or Twitter. This kind of responsiveness to its surroundings is suggestive of understanding of some sort.[26]

Ravenscroft offers a response to the view that entities other than human beings cannot be minded, namely that these kinds of arguments are guilty of chauvinism – that because we are accustomed to think of minds as something that only human beings have, or at least that minds are things that are instantiated by single brains, housed within single skulls, we balk at the idea that there could be minds that are very different from our own (Ravenscroft, 2005). Eric Schwitzgebel, in a paper explicitly about the possibility of groups being conscious, echoes this view, and even gives a name to the chauvinism directed toward group entities in this respect, saying, "It's hard to imagine a prejudice more deeply ingrained than our prejudice against entities that are visibly spatially discontinuous […] let's call such a prejudice contiguism" (Schwitzgebel, 2015, p1699). As jarring as it may at first seem, if an institution is functionally equivalent to a human being, responding appropriately and avowing beliefs, then we can ascribe certain mental states to that institution. After all, according to functionalism, beliefs would be the kinds of things that could be instantiated by a great number of different realisers. If a large institution like the Chinese state could instantiate some or other belief state that was functionally equivalent to some or other belief state that an ordinary individual person had, then refusing to attribute the belief to the Chinese state may well amount to simply not taking functionalism seriously enough.

Yet one might insist that, even so, the Chinese state or any other institution may still be unable to experience qualia – something that would appear to be hallmark of a number of mental states. Christian List poses an interesting question in this regard. He notes:

> "There is something it is like to be a human being; you are experiencing it right now. Presumably, there is also something it is like to be a chimpanzee. Similarly, other mammals, such as cats, dogs, and bats, are plausibly conscious: they experience the world, have perceptions, feel pleasure and pain, even if we cannot appreciate what it is like to be in their position. By contrast, rocks, tables, and chairs lack consciousness. There is nothing it is like

[26] Indeed, similar situations form the basis of our attributions of understanding and various other mental states to other human beings.

> to be such an entity. So, are group agents more like cats, dogs, and bats in this respect, or more like rocks, tables, and chairs?" (List, 2015, p1 [working paper])

Nevertheless, even if critics are correct and there is nothing it is like to be the Chinese state,[27] the Chinese state may still be able to manifest *some* mental states – after all, some have argued that certain mental states need not have any qualitative aspect to them:

> [B]eliefs do not have qualitative character, or need not. What is it like to believe there is a speaker in front of you, or a chair beneath you, if we subtract the visual, auditory, and tactile experience? Nothing, so far as qualitative character is concerned. (Davis, 1982, p196)

Thus, even if functionalism cannot account for phenomenological experience, it still provides us with a good account of mental states that need have no phenomenological aspect. Consequently, even without an ability to experience qualia, it's possible that the Chinese state or other institutions could plausibly instantiate certain mental states like beliefs.[28]

In a slightly different vein, Andy Clark and David Chalmers argue in their paper, 'The Extended Mind' (1998), that the mind, or cognitive processes, needn't be bound by the demarcations "of skin and skull" (Clark & Chalmers, 1998, p7). Once we accept functionalism, then there is reason to accept extended minds. They argue explicitly for the possibility of certain external features of someone's environment that are intimately linked to the cognitive process of that person as legitimately being recognised as part of the whole cognitive system (i.e., that a person and a notebook can count as being part of one cognitive system). Yet their suggestion can be extended to seeing groups of people as forming a single cognitive system. Indeed, at the end of their paper they gesture towards this possibility by noting that "if the view is taken seriously, certain forms of social activity might be reconceived as less akin to communication and action, and as more akin to thought" (Clark &

[27] While some do argue that certain groups may be capable of certain feelings or emotions (see Gilbert, 2000), or of being conscious (see Schwitzgebel, 2015), I will not be pursuing such claims here.

[28] Interestingly, in response to the thought that entities like groups cannot feel emotions like remorse, "because remorse is an emotion, and groups cannot have emotions of their own. Their individual human members can, of course, have emotions of their own, but groups cannot" (Gilbert, 2000, p125), Gilbert suggests that entities like groups are able to feel remorse as a group in the following manner: "A group G feels remorse over an act A if and only if the members of G are jointly committed to feeling remorse as a body over act A […] the joint commitment of the parties is (more fully) to constitute as far as possible by their several actions, including utterances, a single body that feels remorse" (2000, p135). More needs to be said regarding exactly what Gilbert means by joint commitment, but I would argue that Gilbert is clearly offering a very functional account of remorse here – although again, whether or not the group has any phenomenal experience of what it is like to feel remorseful remains debatable. Gilbert herself suggests that it is possible to genuinely feel remorse while not experiencing any of what she calls "feeling-sensations", or "pangs of remorse" (2000, p135) – so perhaps a much broader range of mental states than one would anticipate needn't necessarily have any peculiar phenomenological character.

Chalmers, 1998, p18), thus opening up the possibility of talking about minds that could be made up of more than one person.

Clark and Chalmers suggest that one might criticise their view by suggesting that cognitive processing might be assisted by external means, but that genuine mental states (i.e., beliefs, desires and emotions etc.) are still bounded by the brain (1998, p12). However, using the example of Otto (who suffers from Alzheimer's) and his notebook, they suggest that something external, like a notebook, can play the same role in cognition as something that is normally internal (stored in the brain), namely the memory (1998, p12-14). This is something functionalists would surely be happy with. They suggest that when Otto consults his notebook, what is written there can be regarded as his beliefs, beliefs which he now has access to in exactly the same way as someone who recalls information via the process of remembering would have access to what they believe. They suggest that the functional role played by the notebook in Otto's case, and someone's recalling a belief stored in their memory in a typical case, are exactly the same. They argue that it would be odd to suggest that Otto didn't believe *p*, but someone else did believe *p*, simply because one of them was able to recall something in their head as opposed to making use of a notebook for recalling that information. Chalmers and Clark endorse a functionalist framework when they note that, "[w]hat makes some information count as a belief is the role it plays, and there is no reason why the relevant role can be played only from inside the body" (Chalmers & Clark, 1998, p14). And, as has been suggested, other people – a secretary, or a personal assistant, say – could serve the same function as Otto's notebook in the previous case, and thus a cognitive system, capable of belief, could be made up of more than one person.

However, one may worry that the most that could be said is that other people could be part of Otto's mind, and not that Otto and the other person could form something like a 'group mind'.[29] However, if one person wasn't reliant on another to fulfil their own cognitive functions, but rather if both people were involved in the cognitive processes related to agreed upon concerns, then perhaps together those people could form a joint thinking subject. When it comes to institutional minds, we can imagine, for example, that the notebook does not belong to Otto, but is rather a communal one where a group of individuals involved in the same research project each write their own findings. Otto may have a specific role within the

[29] To clarify, if another person is assisting Otto in remembering dates and addresses that are important to him personally, then they could simply be filling in for Otto's defunct memory and part of his cognition. However, if the two of them formed an institution, then I would argue that they would give rise to mental states that belonged to said institution.

group, to research a particular topic, just as other members of the group have their own roles. The lead researcher may then refer to what everyone has written in the communal notebook to put together a final report. In this way, the view of the research group as expressed in the final report may differ from Otto's personal views and beliefs about the subject. It becomes difficult to say, then, that the beliefs expressed belong to Otto, or to any one of the other researchers. When it comes to institutional minds in particular, it is not clear that the individuals involved will have any primacy. An employee may 'join' a mind in which she is fungible, and thus the mind of an institution does not depend upon any particular person instantiating it. From the perspective of Clark and Chalmers, if Otto and his notebook – or Otto and his fellow researchers – could be replaced by other people and objects while maintaining the cognitive system, one might have further reason to see the cognitive system as something separate from the particular individuals involved.

Section 3: Institutions and The Intentional Stance

When it comes to attributing mental states to institutions rather than individuals, Tollefsen suggests that we adopt what she calls an interpretivist approach, an approach which follows the adoption of Daniel Dennett's intentional stance, and takes our practice of attributing mental states to institutions seriously. She suggests that in attributing mental states like beliefs and desires to an institution, we can successfully explain and predict its behaviour in light of these.

For Dennett, when puzzling over how to explain why something, or someone does what it does, there are a number of different explanatory models one can adopt. He suggests that there are (at least)[30] three positions or stances we could take towards an entity when trying to explain its behaviour, or predict how that entity will behave – namely the physical stance, the design stance, and the intentional stance. When adopting the physical stance, we predict the behaviour of something based purely on its physical make-up and our knowledge of the laws of nature (Dennett, 1971, p88) – chemical reactions, for example, could be predicted from such a stance. Dennett proposes that this is also the stance that we often have to adopt when something we have designed malfunctions (Dennett, 1971, p88-89) – e.g., when my watch stops I may have to open it up and look for physical defects. Alternatively,

[30] In 'The Conditions of Personhood' Dennett also talks of the possibility of a "personal stance" (1988, p148) – an attitude taken towards persons. For further discussion of this point, see Chapter 3.

adopting the design stance entails predicting behaviour based on what one knows about what the system in question was designed to do (Dennett, 1971, p88) – predicting that the time my digital watch face shows will proceed from 13:59 to 14:00 could be predicted from the design stance.

When adopting the intentional stance towards something, or someone, one attempts to explain and predict the behaviour of the entity in question by positing it as an intentional agent with beliefs, desires and intentions (Dennett, 1981, p59), and "by assuming rationality" (Dennett, 1971, p93). For example, in order to make sense of the behaviour of someone who grabs an umbrella before exiting a building, one assumes that they must have a belief that it is raining, or will rain shortly, and that an umbrella will keep them dry, and a desire to remain dry etc. Doing so explains their behaviour – grabbing an umbrella, given these beliefs and desires, is the rational thing to do. Moreover, in observing for oneself that it is raining, one should be able to predict the kind of behaviour just described from those around one by adopting the intentional stance towards them, by attributing to them a certain set of beliefs and desires one thinks a rational agent would have in this situation.

Similarly, hearkening back to the headline: "Spy poisoning: Russia expels 60 US diplomats in tit-for-tat measure", if we see Russia taking the expulsion of Russian diplomats from various countries (including the US) following the poisoning of a former Russian agent and his daughter in the UK as an insult, this allows us to make sense of its behaviour by seeing it as a "tit for tat" measure that Russia undertook. In order to make sense of why Russia did this one must presume that Russia believes itself to be innocent of the poisoning (or, at the very least, has a belief that purporting its innocence is the most expedient thing to do), among many other beliefs and desires, etc. It is also possible that one could predict that this would be Russia's reaction if one adopted the intentional stance towards Russia and assumed such beliefs and desires on its part.

Dennett notes that we can choose to predict the behaviour of a single entity from any one of these three stances – and indeed that it may be possible to successfully predict its behaviour from *all* three of these stances – but that one or another of these stances will likely lead to greater success, or at least, be more easily or readily applicable depending on the entity in question that one is trying to make sense of. He cites the example of the supercomputer Deep Blue, and argues that even though it has been put together and carefully designed, its complexity has reached such heights that we may struggle to accurately or easily predict what it will do if we look at it from the physical, or even the design, stance. On the other hand, viewing the computer as something like an intelligent person with knowledge

regarding the game of chess, that has the goal of winning the game – in other words, adopting the intentional stance toward Deep Blue – may actually prove to be the most effective way of accurately predicting what it will do (Dennett, 1971, p89).

One may worry that, while the intentional stance may well be the best stance to adopt when trying to predict how certain entities will behave (like chess-playing supercomputers, or other people), it is also the case that we may be able to adopt the intentional stance with reasonable success even when it would be inappropriate to do so, for example, when attempting to explain why a thermostat registered a change in temperature (i.e., by citing its belief that it had gotten colder, and its desire to let others know). Thus one can technically adopt the intentional stance toward something that is not a plausible candidate intentional states. Nevertheless, this needn't pose too much of a problem. As the behaviour of the thermostat is explainable from other stances (and perhaps more readily explainable), this should give us good reason to avoid positing it as an intentional agent with beliefs, desires and intentions.

When it comes to institutions, while it may seem far too complicated to try to explain or predict behaviour at a physical level, one may be tempted to adopt the design stance towards them. Institutions are, after all, artifacts. They are something that we have created for a purpose, and which we design explicitly to meet that purpose. Yet institutions often have complex internal structures, including a great number of members occupying different, specialised roles, interacting in a complex hierarchical network with one another, and decision-making structures which may be implemented in a number of ways etc. We may know that a motor vehicle manufacturer has been set up and designed to manufacture motor vehicles, but this does not necessarily make it easy to accurately predict its behaviour. Consequently, while it might, technically, be possible to explain the behaviour of an institution from the design stance, if one were to adopt the intentional stance towards it and viewed it as an intentional agent we might be in a better position to predict its behaviour.

Tollefsen herself suggests that we can be successful in interpreting and predicting the behaviour of institutions when we adopt the intentional stance towards them – making a similar claim to Dennett when she emphasises that, "if a system exhibits complex behaviour patterns that can be successfully interpreted from the intentional stance, then we have every reason to believe that we are dealing with an intentional agent" (Tollefsen, 2015, p110). As noted above, we certainly do talk of institutions as if they had beliefs and desires, and this in turn does seem to play a role in how we understand the behaviour of institutions like governments and corporations within our societies. Moreover, these sorts of attributions play

a role in predictions we make about the behaviour of such institutions. Tollefsen states further that it is not just the case that we can technically adopt the intentional stance towards institutions (perhaps like we can technically adopt it towards inanimate objects like thermostats), but rather that "doing so is explanatorily powerful" (Tollefsen, 2015, p102). When it comes to the thermostat, however, we do not need to adopt the intentional stance, there are perfectly workable alternatives – indeed, "[t]he physical or design stance is more powerful" (Tollefsen, 2015, p102). Yet when it comes to institutions, adopting the intentional stance towards them is, more often than not, more powerful than adopting either of the other two stances towards them. When one makes use of group mind talk, attributing mental states to institutions and viewing them as agents, one is often in a better position to understand and explain what is really going on in the social world.

To illustrate this, Tollefsen extends Dennett's thought experiment regarding the intentional stance and a prediction contest to group mind talk. She imagines that two people are engaged in a prediction competition in which one person can only make reference to individualistic explanations (like those one would appeal to from a design or physical stance), while the other person, adopting the intentional stance, is able to view the organisation as a rational agent (with beliefs and desires etc.), and makes a prediction on that basis (p106). In her particular example the two people have to predict what Ford will do in response to an enormous increase in fuel prices (p106). When it comes to predicting what Ford will do, Tollefsen notes that the person who is the individualist will have to find out who makes the decisions in the company, and perhaps find out how they voted in the boardroom, and on what propositions, as well as what mechanisms are in place within the company, what infrastructure is available and how that's all used etc. in order to find out and explain what Ford will do. On the other hand, the person who is viewing Ford as an intentional agent will be able to predict quite easily, by positing what mental states Ford would be likely to have, that because Ford has lots of large vehicles that are fuel heavy that it wants to be able to sell, and because Ford believes consumers are unlikely to want fuel-heavy vehicles given the enormous increase in the price of fuel, Ford will offer a discount on these vehicles in the near future. While Tollefsen herself suggests that someone who was committed to explaining things in individualistic terms may well be able to come up with the same prediction (just after a much lengthier and more difficult investigation) she also notes that, compared to the methodological individualist's own efforts, "the collectivist's performance will look like magic" (Tollefsen, 2015, p106). Tollefsen does note that the person who adopts the intentional stance towards the institution may not always make an accurate prediction, that

their prediction could be a bit "risky" (p106) – yet in much the same way that we make inaccurate predictions about the behaviour of other individuals on occasion, a prediction from intentional stance will likely still be a "pretty good one" (p106).

Dennett claims in his paper on the intentional stance that his "thesis will be that while belief is a perfectly objective phenomenon […], it can be discerned only from the point of view of one who adopts a certain predictive strategy, and its existence can be confirmed only by an assessment of the success of that strategy" (Dennett, 1981, p59). Dennett calls himself a realist, and thinks that there is some fact of the matter about whether or not a person, or indeed an institution, believes *p*, while noting that we only have (imperfect) access to that fact via observational means – i.e., by observing how a person or institution behaves, and interpreting that behaviour. Just as functionalism opened up the possibility of group mind, this is a position which allows for the possibility of group mental states in light of the promise of success in adopting the intentional stance towards groups like institutions. The two positions are consistent with one another, as the beliefs that one attributes using the intentional stance are arrived at in just the way that a functionalist would attempt to identify particular beliefs (by looking at environmental inputs, behavioural outputs, etc.).

One might be inclined to object again that the suggestion that institutions could have mental states seems to be a violation of Occam's razor in that it involves positing a whole bunch of extra and unnecessary mental states in the world. This is because "the individual-level explanation makes reference to the cognitive states of individual employees. The group-level explanation makes reference to these cognitive states plus an extra set of cognitive states at group level" (Sprevak & Statham, 2015, p10). If one is able to offer an adequate explanation in purely individualistic terms, then this is what we ought to do – such an explanation would clearly be the simpler one. On the other hand, Sprevak and Statham, like Tollefsen, suggest that when it comes to explaining the behaviour of institutions, if one is able to refer to the institution's mental states this is so much more straightforward, and so "if a group-level explanation […] is available, many stages of the individual-level explanation seem pointlessly complex" (Sprevak & Statham, 2015, p15). In other words, one of the great strengths of a position like Tollefsen's is that she can give a simpler, more ecumenical explanation by utilising group mind talk. Doing so can lead to better predictions too.

There is however also a stronger claim that can be made here, and that is that there will be situations in which it may well be practically impossible that we could actually predict what the views or actions of an institution would be, or see what was really going on, if we kept to the individual level. In part I think this is true because it seems likely that it will

be practically impossible for us to know what every relevant individual within an institution thinks and intends, and the precise mechanisms by which these thoughts and intentions relate to one another. It is not enough to know that a particular employee at Ford is thinking about different font colours – one would need to know what a number of other relevant employees' mental states were, what roles these employees occupied, and how they were all related to one another in the institutions hierarchical structure, etc. – in order to be able to glean that Ford was engaged in advertising a new car, for example. Accessing this kind of information could be extremely difficult and time-consuming. Moreover, it is not just the case that we often know very little about the individuals who make up an institution and how they relate to one another, there are other situations in which it will be extremely difficult for us to know exactly what is going on at an individual and internal level because of secrecy – the workings governments and certain government agencies, for example, often remain opaque to the public for a number of reasons, including security concerns or even just one-party rule.

Yet even if one did have access to the mental states of all relevant parties within an institution, it may well be the case that one would not be able to work out what the institution itself was doing or saying if one was entirely focused on viewing the world in individualistic terms. For example, we can imagine Ford Motor Company placing an order for certain goods with another company via the phone, and as Dennett notes, if someone has failed to see that "indefinitely many different patterns of finger motions and vocal cord vibrations – even motions of indefinitely many different individuals – could have been substituted for the actual particulars without perturbing the subsequent operation of the market, then they have failed to see a real pattern in the world they are observing" (Dennett, 1981, p69). Tollefsen herself notes that "the interpretive stance we take toward groups is able to discern real patterns of social behaviour, patterns that are missed if one attempts to explain the social world by appealing only to individual intentional states" (Tollefsen, 2015, p106). This is because actions, like mental states, can be multiply realisable. In an institutional setting, the institution is multiply realisable too with regard to the individuals that make it up.

Accepting that particular group-level mental states could be multiply realisable involves thinking of an institution's mental state as something which supervenes on the particular configurations of the institution. Functionalists often appeal to the idea of supervenience to explain the relationship between the base that realises the mental state, and the mental state itself. They suggest that while you can't have a difference in the property (the mental state) that supervenes on an identical base realiser, you can have multiple realisers for the same supervening property (mental state). One may worry that all

supervenience claims about mental states must lead to reductionist talk (Kim, 2000) – and that if an institution's mental state supervenes on the mental states of individuals and the relations between them, there is an individualistic story to tell that trumps all other stories. However, Donald Davidson argues that "dependence or supervenience of this kind does not entail reducibility" (Davidson, 1970, p214). Even though a particular instance of institutional belief may be identical to some complex set of individual beliefs and their inter-relations etc., given the fact that the exact same belief or action can be instantiated in different ways by different groups, retaining talk of beliefs, intentions and actions at the institutional level becomes indispensable.

To elaborate on the difficulties involved, let us return to the example of advertisements. When trying to explain that what an institution has done is advertise its goods or services, if one did not talk of the aims/desires/intentions of the company as a whole, one might genuinely miss out on the fact that it had advertised anything at all, even if one knew what the individual intentions and beliefs of all the members of the company were at the time that they advertised their product. This can be the case because when an institution advertises their products or services they can do so in a number of wildly varying ways – from handing out free samples in a supermarket, to dressing up in costumes and handing out flyers, to putting up posters, to promoting goods in a newspaper, to putting a commercial on television, posting something on social media, or setting up banner advertisements on certain websites, etc. – and in addition they can do so for various reasons – to generally promote their product, to alert the public to a sale, or to get rid of goods that are approaching their sell-by date, or even, perhaps surprisingly, to apologise.[31] It can also be further complicated because the institutions could outsource (some of) their advertising, leaving it up to someone external to the institution.

It is useful to be able to say that two different companies (with different individual members, with different organisational structures, and for different reasons etc.) have done the same thing (e.g., advertised their product, recalled a product, gone on strike, celebrated a five-year anniversary, or apologised etc.) in order to be able to make sense of, and successfully function in, the social world. Yet if one looked at each of the companies in terms of their individual members and their mental states and interactions, it is possible that the

[31] For example, in order to apologise to customers for running out of chicken and being forced to temporarily close a number of its restaurant chains in 2018, the fast food chain KFC took out a full-paged advert in UK newspapers featuring an image of an empty KFC bucket with the letters of the logo on the bucket switched from 'KFC' to 'FCK'. In an article entitled 'KFC's apology for running out of chicken is pretty cheeky', it was noted that "The ad has gone down well on social media - we Brits appear to be a forgiving bunch" (BBC News, 2018).

activities and underlying structures of these companies could look completely different. As noted, at the very least because they would probably be made up of different individuals, but also because different companies could reach the same decision to do *x* in an indefinite number of ways, or supervening on an indefinite number of configurations of individual mental states. From here it becomes unclear how one would be able to say with ease that the companies had done the same thing after all unless one reverted to group-level explanations.

If one considers examples of corporations advertising or recalling products, it seems that identifying mental states or actions at the institutional level is what allows one to say that two completely different institutions, made up of different people and organised in different ways, have done, or believe, the same thing. As Tollefsen notes regarding possible explanations of institutional behaviour: "to answer these questions we cannot appeal to individual intentional states because, in many cases, there may be nothing in common at the micro-level that would explain why these companies […] acted in the same way" (Tollefsen, 2015, p88). Both KFC and Ford could have advertised their product, the same action – however one did so by paying volunteers to wear drumstick suits outside their restaurant chains, and one did so by having their advertising team put together a printed advert to be published in a newspaper, with both institutions' members experiencing greatly differing mental states during these respective processes from one another. Perhaps if there were employees in both companies that had the intention 'to advertise', or 'for the company to advertise' when performing their actions, one would be able to determine on an individualistic level that this is what both companies were doing. Yet in many institutions tasks will be broken down between different sectors to such an extent that individual employees will have intentions to do with their own specific task (like 'write a blurb', or 'print something out') rather than intentions reflecting the overarching aim of the institution as a whole. In fact, as will be discussed in Chapter 3, Gilbert (2000) and List and Pettit (2011) even suggest that groups can have mental states that no individual member of that group has – which means that it is entirely possible that if one interrogated the mental states of all employees in a company, no one in the company may have the intention to advertise, and yet the company could still intend to advertise. This is also true with regard to the action of advertising. No one person is advertising, the company is.

Yet it could still be argued that the success of the intentional stance is again not strictly speaking a thoroughly convincing argument for the existence of group-level mental states and agency. In fact, Dennett remarks on the power of the intentional stance by saying that it can be successfully applied to almost anything (animals, artefacts and plants in

addition to persons), and that "even in the worst of these cases, even when we are surest that the strategy works for the wrong reasons, it is nevertheless true that it does work, at least a little bit" (Dennett, 1981, p66). We have seen that the physical stance is unlikely to be successful when applied to understanding the behaviour of institutions, and that although it may be possible for the design stance to be successful on occasion when applied to institutions, its success is dim when compared to that of positing institutional intentional states. Although adopting the intentional stance towards institutions can be successful (just as it is with people), perhaps more needs to be said to ensure that this practice is working for the *right* reasons.

Thus, despite all of the above, it may be argued that it remains an open question whether or not a methodological collectivist (as opposed to individualist) approach is a purely pragmatic or useful tool, or whether an institution really has mental states of its own. The success that can readily be had by ascribing mental states to institutions – that this practice holds the promise of such great explanatory and predictive power – is surely a reason to be sceptical of the argument that we ought only be concerned with reducing group phenomena to an individual level in order to enjoy any success in this field, and perhaps this is enough for some. Nevertheless, there is another concern at hand – whether or not institutions are genuinely so minded affects attributions of accountability. In adopting the intentional stance towards institutions, we are offering explanations (that may be right, or wrong) for the behaviour of those institutions. Part of what allows us to form judgements is that we have explanations for their acting as they did at hand. That our explanations are correct would seem to be of some importance in determining how appropriate whatever our judgements may be in this regard. With this in mind I would now like to look at more explicit arguments for the existence of institutional mental states in Chapter 3.

Chapter 3: Institutions can be Minded and Moral Agents

> "Tiger Brands Refuses to Take Blame for Listeriosis Deaths" - Pelane Phakgadi, *Eyewitness News* (2018)

In Chapter 2 we saw that functionalism about the mind is consistent with groups like institutions being minded. In addition, if we adopt the intentional stance towards institutions we are often able to successfully make sense of what institutions do, and to make good predictions regarding their behaviour. Nevertheless, the possibility that certain institutions could instantiate minds does not mean that they in fact do, and the apparent success of adopting the intentional stance towards institutions does not guarantee that institutions are in fact minded.

In this chapter I now turn to the arguments of Margaret Gilbert, and Christian List and Philip Pettit, that deal explicitly with existence of group-level mental states (in Sections 1 and 2). I will then look at List and Pettit's further arguments for seeing institutions as agents in Section 3. Finally, in Section 4 I discuss suggestions that have been put forward for regarding institutions not only as agents, but as agents that meet the criteria for being appropriate bearers of moral responsibility. Some regard the latter as the hallmark of personhood, something which distinguishes human beings from animals (Menkiti, 1984; 2004). Despite an instinctive reluctance (at least on my own part) to considering institutions as persons, ultimately, I find no compelling reason to exclude institutions from being categorised as such. Regardless of one's views on attributions of actual personhood to institutions, what is important is that we can rightfully hold institutions morally accountable. Once we acknowledge that we can do so, we can begin to seriously look at the unethical behaviour of institutions with an eye to understanding and preventing such behaviour, or perhaps meting out punishment where appropriate.

Section 1: Gilbert and Group Belief

According to Gilbert, a group can have a belief, *qua* group. She defines a group belief as follows: "there is a collective belief that *p* if some persons are jointly committed to believe as

a body that p" (Gilbert, 2000, p39).[32] She emphasises that a joint commitment does not merely amount to a number of individuals having commitments that coincide, or commitments that are simply suitably related to one another, or commitments that two or more people have in common. Gilbert is in opposition to individualistic views like those of Michael Bratman, who says "I see shared intention as consisting in relevant intentions of each of the individual participants, in a suitable context and suitably interrelated" [Bratman, 2009, p42]). For Bratman, shared intentions are not the irreducible intentions of the group and do not require any reference to something like a group mind – shared intentions arise out of the intentions of individuals in a suitable setting. Gilbert, on the other hand, argues that a joint commitment to *x* is formed by each individual group member giving a kind of conditional commitment to *x* as a body, in much the same way as an individual would *x*. This conditional commitment depends on other's similar conditional commitments, such that no one is committed until everyone is committed. For example, one may imagine two friends, Maya and John, who want to go and see a movie, and neither wants to go to the movie alone. Maya may say 'I'll go to the movie if you do', and likewise John may say something like 'well, I'll go if you will'. Thus each manages to give a kind of conditional commitment to going to the movie, such that if either of them had said they wouldn't go, the person who had offered a conditional commitment to the activity wouldn't actually be committed to going.

While the examples that Gilbert refers to tend to be those of smaller or more informal groups, one can nevertheless imagine similar scenarios arising in larger institutions – for example where certain sectors of an institution are in charge of a particular task, but require the approval of the CEO before going ahead with their proposed course of action. The CEO may not have any expertise in the area, and so give conditional commitment in the form, "Let's do it if you're happy to go ahead," while those charged with the task will offer their conditional commitment along the lines of "well, we'll go ahead with it if you're happy for us to", etc. In this way institutions can come to jointly commit to perform certain acts, or hold certain beliefs, such as "this is the best course of action," because they have jointly committed to believe that that is the best course of action, by each relevant group member offering their conditional commitment to the act or belief. The founding documents of an institution may also require members to jointly commit to certain beliefs, for example, by

[32] With Gilbert's suggestion that 'some persons' could give rise to a collective belief, one may be inclined to think that what she is saying may not apply to institutions. However, as Gilbert appears to have far less stringent requirements for a group of people to form what she refers to as a plural subject and so become the appropriate subject of a collective belief, this need not be a problem for the suggestion that institutions are also capable of meeting such requirements.

stating that 'those who join the institution mutually agree to promote *a* and *b*', where this generates the institutional beliefs that *a* and *b* are valuable etc. Moreover, an institution itself is likely the product of joint commitment. When one focuses on institutions as organisations or societies which are founded for a particular purpose, joint commitment between some members seems essential to the establishment of an institution, and one can even view an employment contract as a kind of mutual conditional commitment.[33]

Additionally, no one individual has the power to rescind on the joint commitment, contrary to what one would expect if it were the case that these were individual, personal commitments (Gilbert, 2000, p148). As Gilbert points out, if an individual who is jointly committed to some end does choose to rescind on a joint commitment, they will justifiably be subject to censure from the rest of the group (2000, p148). For example, we can imagine that a philosophy department needs to decide what their policy regarding accepting late assignments should be. After meeting with one another, it is decided that the department will not accept late assignments. If a member of the department then goes on to accept late assignments for their course, they could rightly be criticised by other members of the department because it had been agreed that the department would not accept any late assignments. Of course, this doesn't entail that a joint commitment cannot, or shouldn't, be rescinded at all. Such an overturning can take place, but the group as a whole must agree to rescind the joint commitment. Thus a member of the department could make an appeal to the rest of the department regarding allowing late assignments to be submitted (perhaps under special circumstances), but they cannot simply choose to accept late assignments of their own accord – not without committing some sort of violation for which they would face censure. Due to the fact that individual group members cannot just unilaterally opt out of such an agreement without facing censure, joint commitments cannot be understood on a purely individual level (2000, p148). This, in essence, is Gilbert's account of how group action is to be properly understood – through the adoption of a joint commitment to *x*.

Such an approach also structures her account of group belief. Thus, to spell out her account of group belief more explicitly, Gilbert explains that the joint commitment to believe *p* entails that:

> those concerned confidently express the view that *p* in appropriate contexts and do not call it or obvious corollaries into question. Their behaviour generally should be suggestive of a

[33] It may be possible for a single individual to register a company, bringing that institution into existence. Yet as soon as other individuals join that company, they will need to jointly commit to a number of core ideals or aims of that institution.

> belief that *p* in the appropriate contexts: if we believe that *p* as a body, the actions of each of us should reflect that fact. (Gilbert, 2000, p41)

With this in mind one can see that Gilbert's view can also be understood within a functionalist framework – a particular mental state, like a belief, is something that can be understood as having a certain functional role. Institutional beliefs represent the way the world is as the result of certain inputs, and so serve to guide the institution's actions in the light of its desires and intentions, thus causing certain outputs. On Gilbert's picture, a philosophy department's belief that *p* (that the best policy to have with regard to late assignments is not accept them) is a belief in virtue of playing a certain role – i.e., affecting behaviour of the institution in a very particular way. It can be brought about by certain information with respect to the submission of late assignments being brought to the department's attention (e.g., by witnessing or hearing how other departments with more lenient policies fare), and by having members of the department sign contracts detailing certain attitudes (e.g., when accepting a job at the university they have to agree to follow certain procedures), or by having members of the department agree to abide by the outcome of certain decision-making procedures (e.g., by going with what the majority of the department think), etc. Their having the belief that *p* then affects certain other beliefs that they may hold (e.g., the members of the department do not call obvious corollaries to *p* into question), and leads to certain behavioural outputs (e.g., the members of the department confidently express the view that *p*, and act as if *p* were true). And none of this requires that the institution itself experience any qualia, something which is open to doubt. As noted previously (see Chapter 2), beliefs do not necessarily have a qualitative feel to them, and Gilbert offers an account of a group believing something that doesn't require it to experience qualia of any kind.[34]

[34] As noted above (see Chapter 2, footnote 28), Gilbert does suggest that groups are capable of other kinds of mental states that we ordinarily think of as having a certain qualitative character i.e., her account of 'collective remorse' (2000). Yet arguably, given the aforementioned concerns regarding certain mental states essentially requiring a certain qualitative feel, it's not clear that groups *could* feel remorseful as groups. However, one may be able to make sense of this if one sees that remorse also seems to necessarily involve holding beliefs (i.e., the belief that I, or that the institution, acted wrongly). Importantly, Gilbert thinks that accounting for group remorse in terms of individual members feeling remorseful is simply not good enough - even if members of a group genuinely express remorse over a group act (and not merely personal remorse over their own contributions to the group act). This is because the result would simply be "a number of separate subjects of remorse rather than one undivided subject of remorse – the group" (2000, p.134). Thus, what she takes to be necessary for a group to be capable of remorse (something that our everyday way of speaking seems to suggest is possible, and something that is often called for in terms of certain groups being asked to apologise for historical acts like slavery etc.), is for the members of a group to be jointly committed to feeling remorse, as a body, over a particular act (2000, p.135). Again, perhaps to make sense of the suggestion of institutions or other groups

Gilbert insists that such a belief should properly be understood as the belief of the institution as a whole, and not the belief of any, or any number of, group members, because the institution's belief can be at variance with what its individual members believe. Gilbert thinks that this notion is already deeply embedded in our way of understanding the social world. Thus she notes that when a member of an institution disagrees with the institution's view, by saying something to the effect of "personally speaking" (2000, p41), that member can put forward their individual view. Gilbert further asserts that "the availability of such locutions strongly suggests that a group's belief is not understood to be a belief common to all of the group's members; more generally, it strongly suggests that one's personal view is conceived as distinct from the view of a group of which one is a member" (2000, p41). Gilbert also claims that it is entirely possible for a group like an institution to jointly commit to believe *p* even if not a single member of that institution actually believes *p* (1987, p195-197). Nevertheless, even though they personally do not believe *p*, a person who is a member of an institution would also be perfectly justified in saying 'we believe *p*', given their participation in a joint commitment to believe *p*. Given that one may have to commit to certain institutional beliefs when one joins an institution, this may be a fairly frequent occurrence.

Gilbert criticises accounts that see collective belief as a mere sum of the beliefs of the institution's members, as this amounts to merely a number of distinct subjects that happen to believe the same thing. This is not sufficient since, as Gilbert points out, it is entirely possible for everyone in some or other institution to believe *p*, but be unaware of the fact that everyone else also believed *p*, and perhaps even be afraid to admit their own belief that *p* to the rest of those in the institution (1987, p187). For example, I, a member of a philosophy department, may secretly be a Flat-Earther, and for whatever reason ashamed to admit it. Unbeknownst to me, the majority of members of my philosophy department are also Flat-Earthers, but have also kept their beliefs to themselves. Moreover, the philosophy department has never had cause to take a public stance regarding whether or not the earth is flat – the issue has simply never arisen. Gilbert suggests that, in this situation, it would be extremely odd to say that the philosophy department believes that the earth is flat, and thus that "it is not logically sufficient for a group belief that *p* that most members of the group believe that *p*" (Gilbert, 1987, p187). Gilbert argues further that, even if it were common knowledge within

'feeling' remorse, one would have to substitute for such feelings beliefs that they had acted wrongly, and perhaps a wish or intention to make things right.

the institution that everyone individually believed *p* as all members believed *p* and were very vocal about it, this still would not be enough to constitute an institutional belief. Indeed, we can imagine any institution, and accept that all or most of the members of that institution probably believe that the earth is round, without this actually constituting what we would call the institution's belief as opposed to say a common or generally held belief (Gilbert, 1987, p188-9). Thus, for Gilbert it is essential that members of an institution are jointly committed, as a body, to believe *p* in order to attribute that belief to the institution and differentiate institutional beliefs from the beliefs that members happen to have in common. Without some or other agreement between members of the institution to espouse a certain view together, we are left with a situation in which "the fact that a group is involved does not play any obviously essential role in what is going on" (Gilbert, 1987, p189).

Yet one can imagine a situation in which a certain belief that *p* is not spoken about, but simply assumed by all members of an institution, and thus that the institution will act as if *p* is true. Even though the institution lacks official or explicit awareness of the belief, and certainly lacks a joint commitment to believe that *p*, this could nevertheless be a belief that we would be justified in ascribing to the institution. Indeed, if one were to adopt the intentional stance toward said institution one may well have to attribute the belief that *p* in order to successfully make sense of its behaviour. So even though the institution lacks official or explicit awareness of the belief, and certainly a joint commitment to believe *p*, this could nevertheless be a belief that we would be justified in ascribing to the institution. If this were the case, then there may well be some situations in which a simple majority will be sufficient for institutional belief. Yet it would seem to be the case that we would only be able to attribute a belief to an institution on this basis if this were a belief that informed the behaviour of said institution, like the belief that the best policy to have is one where students are not allowed to hand in their assignments late – and not in cases where the fact that the majority of members believe a particular thing, say that the earth is flat, but where this has no bearing on the behaviour of the institution in question. With this in mind, one might think that just as a simple majority is not necessary for group belief, so too is joint commitment not necessary for group belief. If we can rightly attribute a belief to an institution on the basis that this belief is what is informing its behaviour – a belief which members did not jointly to commit to believe – then Gilbert's requirements for a group belief do not necessarily hold in all situations. Yet Gilbert may still be able to claim that if this belief that has not been explicitly committed to were drawn to the attention of the group, they may well explicitly

endorse it. It may also be the case that the belief doing the work is one which has arisen as an implication of a range of other explicitly jointly committed to beliefs.

In any case, I think Gilbert's account of group belief is a powerful one. Her account of collective belief captures the sense in which, in certain situations, what an institution believes cannot be straightforwardly understood, or understood at all, in terms of what the individuals who make up that institution believe – and if one accepts her account, one ought to be able to say that a group can form a mind of its own (to borrow a phrase from List and Pettit, both of whom are also concerned with distinguishing a group's beliefs from those of its members). However, one may worry about the degree to which her account appears to make institutional beliefs something that can be under an institution's control – and thus her account takes away from the nonvolitional character of belief, something which many see as an important hallmark of belief (Cohen, 1989, p369). If one reflects on instances of individual belief, it seems that individuals can't help but take their beliefs to be true – that that's what it is to believe something – and what they take to be true doesn't seem to be a matter of choice. I can't just decide to believe that my car is parked on the moon, for example. But on Gilbert's account it looks as if an institution *can* decide what it is that it will believe and, as noted, members of the institution can jointly commit to believing something that no individual member believes (i.e., something that each member may take, individually, to be false). This seems quite odd.

Yet Gilbert could argue that the above is somewhat misguided – the institution does not decide what the institution will believe. What the institution ends up believing is what (at least some key) individuals jointly commit to believing as a body. Once they have jointly committed to believe that *p*, the institution forms a plural subject that has a belief that *p*. Moreover, once this collective belief is in place, it does seem as if it functions in exactly the way we think of beliefs as functioning: it informs and influences behaviour and is responsive to the world – after all, all members of that institution are now committed to behave in a way that is suggestive of a belief that *p* in the appropriate contexts etc. Believing that *p* also entails that the institution not endorse beliefs that obviously conflict with *p*, at least not without revisiting their initial belief that *p*. Gilbert provides us with an example of a poetry discussion group, where the group concludes that the final line of a Philip Larkin poem was quite moving, even though none of the individual group members, "in their hearts" (1987, p191), believe this. In spite of personally disagreeing with the conclusion of their discussion, Gilbert notes that "[w]hat seems to be crucial is that all or most members decided to let a certain interpretation 'stand' in the context of their discussion" (1987, p191). Thus the group now

believes that the final line of the poem was moving. Gilbert goes on to say that when the group moves on to discuss another poem, should a group member say something that contradicted what they as a group believe, for example that unlike the Larkin poem, the final line of the next poem was quite moving, the group would be obliged to chastise that group member and challenge that view, as this statement would contravene the joint commitment to behave in a manner that was indicative of the belief that the final line of Larkin's poem was moving. As such, thus there is also a clear sense in which, although a group may knowingly commit to believe something that not a single group member individually believes, a group's beliefs will nevertheless not be entirely volitional – and that this would be increasingly the case the longer a group continued to exist, as they would inevitably build up a store of beliefs which in turn entailed all sorts of other beliefs.

Yet there are, perhaps, some other reasons to think that Gilbert's account of collective belief may be inappropriate for attributing beliefs to institutions. In this respect, institutions tend to have a very different sort of structure to the groups that usually operate in Gilbert's thought experiments (i.e., two people taking a walk together, or an informal poetry club). Institutions are often deeply hierarchical in nature. Consequently, while there may be members of that institution who are explicitly involved in decision making processes (CEOs and shareholders etc.), there are also many other members of the institution who have absolutely no say in how things are run, or what the institution ought to believe (e.g., the average office or factory worker). One might feel that it could thus be misleading or inappropriate to ascribe a belief to an institution in which not all members have a say in what policy the institution should adopt, or what the institution thinks, as they didn't *all* jointly commit to believe *p* on Gilbert's account. In fact, a number of employees at Google might be blissfully unaware of what it is that Google even believes, never mind jointly committing to believe *p* or *q* as Google.

I believe this concern can be best met in two ways. Firstly, as CEOs and shareholders do arguably form a decision-making group at the head of the institution, there is still a sense in which one can talk of the institution's beliefs, even if it turns out that the whole organisation cannot be said to hold some or other belief (at the very least, what is going on wouldn't be a case of mere individual belief). Secondly, Gilbert herself suggests that it may also be legitimate to presume a kind of background assent on the part of non-decision-making group members to accept whatever those in charge decide. For example, she notes that often where a government can be said to believe something, that belief will be ascribed to the nation as a whole. She suggests that, perhaps by voting the government in, we may "presume

that the citizens at large have endorsed the idea that what their government as such thinks may be regarded as what they as a nation think" (1987, p209). Thus, even if an employee gives no real thought to the decisions of the executives and CEOs, in acquiescing to be part of the company, they could be seen to have given consent to that decision-making structure (in contracts that they have signed when they joined the company, say). Indeed, Kirk Ludwig has expressly argued that even in cases where a) an individual member disagrees with the beliefs, views or actions of their institution (2014, p94), or b) is ignorant of the actual beliefs, views or actions of their institution (either through inattention, or because they do not have information about them) (2014, p95-96), or c) where they have not expressly endorsed the specific beliefs, views or acts undertaken by the institution (2014, p97), they can nevertheless be said to play a constitutive role in the institution's beliefs, views and acts. He does so by pointing out that, especially when it comes to large groups, like institutions, "proxy agency is pervasive" (Ludwig, 2014, p75). By this he means that an institution will often authorise an individual, or a subgroup (like a department) within the institution to act on behalf of the institution as a whole (although they may also authorise an outsider to act on their behalf as in the case of hiring a lawyer, for example, to represent their interests) where their acts will then count as the acts of the institution itself (Ludwig, 2014, p76). That a proxy is appointed to act on behalf of the institution will require that this falls within the gambit of certain institutional procedures and arrangements. Yet that one becomes a member of an institution will typically require that they "agree to the conditions of membership, which includes an endorsement of the institutional arrangements." (2014, p97). Thus, even if one is not directly involved in the choice of proxy, or even in determining the content of the institutional view or belief on some matter, one can still be part of the agent that does x or believes y:

> [I]n one way or another, group action through proxy agency calls upon every member of the group to contribute. For some the contribution may consist in no more than agreeing to the institutional arrangements, which is a kind of commissioning in general the various role fillers in the institution to fulfill their assigned roles, exercising the powers thereby accorded to them. (Ludwig, 2014, p100).

And this does seem to hold true in many (if not all) institutions where employees are obligated to abide by codes of conduct etc. in order to become a member of that institution. That only a particular member or a select subgroup of an institution, say the CEO or the board of directors, goes on to elect spokespersons or make decisions about certain issues does

not mean that other members of the institution are unrelated to those acts if they have agreed to abide by the institutions structures and practices in joining. Although this kind of joint commitment is easier to identify with a company than with a government. An employee at a company signs a contract. A citizen, on the other hand, may never vote or be otherwise politically active. One could argue that by just remaining in the country one gives a kind of background joint commitment, but for many, leaving is simply not possible. As a result, speaking of the beliefs of a nation in this manner may not be possible. However, one can certainly speak of the government employee, who by taking on certain tasks, commits to serve the governmental institution. Consequently, the kind of collective belief argued for on Gilbert's account does still seem viable.

Section 2: Rationality, Consistency and Group Belief

Just as Gilbert thinks it is possible for a group to hold a belief that no individual member individually holds, List and Pettit are particularly concerned with distinguishing the mental states of an institution from those of its members. In their book *Group Agency*, their focus is on certain kinds of groups which they suggest we can see as agents in their own right. Unlike Gilbert for whom jointly performing some or other action would seem to create a 'plural subject',[35] List and Pettit say that this "amounts to an over-ascription of agency" (2011, p216), and that group agents also need to meet a minimum standard of rationality in order to count as agents. For List and Pettit, group agents will be groups like institutions etc. that persist over time. This is also something that comes with the adoption of the interpretative or intentional stance, as one will be unable to interpret an agent by attributing beliefs and desires if they are not rationally constrained.

Yet in agreement with Gilbert, List and Pettit explain that the urge to see the views of the group agent as the sum of all or some of the views of the individual members of the group agent is highly problematic, even if one attempts to take into account how those views relate to each other within a decision making or organisational context. List and Pettit point out that there is a paradox known as the discursive dilemma – a problem which most groups that have to make a number of decisions over time will inevitably encounter – and which they think can only be resolved by a group "developing a mind of its own" (List & Pettit, 2011, p8).

[35] Gilbert has argued that "to have a joint project is to constitute a group" (1989, p377). Yet for Gilbert two people who are rowing in a boat in a lake, come across a drowning swimmer, and embark on saving them have a joint project (1989, p376).

In an earlier paper, "Groups with Minds of their Own", Pettit notes a certain kind of paradox has recently been identified in jurisprudential literature called the doctrinal paradox – but he suggests that this kind of paradox is actually generalisable, and dubs the generalised paradox 'the discursive dilemma' (Pettit, 2004, p168). The discursive dilemma comes to the fore when an institution has to make a decision regarding some or other proposition, where there are a number of interrelated clauses at issue which are related to final proposition. For example, where an institution needs to decide whether or not it believes '*p and q*', where '*p*' and '*q*' are the related clauses at issue. As we saw above, one of the most straightforward ways of thinking about an institution forming a decision, or coming to espouse a certain attitude or belief as an institution, is via majority voting. However, List and Pettit argue that aggregating the views of the individual members of the institution may not lead to what we could call the institution's view (List & Pettit, 2011, p43-44). Problems can arise because the members of the institution form their own opinion or belief about the individual clauses at issue (i.e., '*p*' and '*q*'), and thus also a view on the final proposition (i.e., the conjunction '*p and q'*). Because of this, the aggregate of the institution's members' views on the individual clauses and the aggregate of the institution's members' view regarding the final proposition may be in conflict.

For example, Pettit asks us to consider a scenario where three individuals (Person 1, Person 2 and Person 3) on a housing board need to work out whether or not a landlord is blameworthy and liable for harm caused to tenant by an exploding heater. To find the landlord blameworthy and liable they need to establish *p:* that the landlord was responsible for looking after the heater, and *q:* that the tenant was harmed by the exploding heater (Pettit, 2014b, p167-8). When looking at just *p*, we can imagine a situation in which Person 1 and Person 2 believe that *p* is true, that the landlord was responsible for looking after the heater, but Person 3 does not. Nevertheless, given that the majority of the board believe *p*, one could say that the board believes *p*. When looking at just *q*, Person 1 also believes that *q* is true, that the tenant was harmed by the exploding heater, and this time Person 3 thinks *q* is true too. Person 2, however, does not think *q* is true. Yet here again, given that the majority of the board believe that *q* is true, one could say that the board believes *q*. So we have a situation in which the board can be said to believe *p*, and the board can be said to believe *q* if one takes a majoritarian stance toward group belief. However, if the board were then to vote on whether or not they believe '*p and q*', only Person 1 will say that she believes '*p and q*' (because Person 3 rejected *p*, and Person 2 rejected *q*). Thus, on the majority view, the board does not believe '*p and q*'. Consequently, as we can see from the table below, if we proceed by

aggregation or majoritarian voting then are left with the rather odd situation in which we would say the board believes *p*, believes *q*, but does not believe '*p and q*'.[36]

Simple discursive dilemma:

	Do they believe: P	**Do they believe: Q**	**Do they believe: P & Q**
Person 1	yes	yes	Yes
Person 2	yes	no	No
Person 3	no	yes	No
Majority View	yes	yes	No

Pettit argues that in order to resolve this tension, groups will have to make decisions either 1) via premise-centred reasoning, where the group members put forward their views regarding just the premises (*p* and *q*), and the view of the group regarding the final proposition is derived from the view of the group on the premises, or 2) via conclusion-centred reasoning, where the group members privately take into consideration the premises at issue, but only vote on the final proposition (Pettit, 2004, p172-175). One might not think that this is a problem if the board just opts for option 2 and has a decision-making process in place where they always just vote on the final proposition under consideration. However, one can easily imagine a situation in which members of the institution in question have to interact with the public – perhaps because they are interviewed about their views by the media – and if asked whether or not the institution just believes *p*, or just believes *q*, the members of the institution may well declare that the institution believes *p* and that it believes *q*, as this would indeed be what the institution believed were '*p*' or '*q*' the final proposition being considered. Yet as the institution has already decided that it does not believe '*p and q*', it would then be regarded as inconsistent or incoherent or even hypocritical by the public (and face the negative consequences that result from being seen as such).

[36] Pettit goes further and shows, like Gilbert, that the board is actually able to reach a decision that, unlike the above example, *none* of the board members individually endorse (2004, p.171). Pettit illustrates this with an example of three group members voting on three inter-related premises, and a final proposition. All three board members endorse two of the three premises, in such a way that each member disagrees with a different premise. Thus, when one determines what the board thinks about each premise, one will end up with a majority endorsement. However, given that endorsing the final conclusion depends on endorsing all three of the premises, while the board will endorse the final conclusion, no individual group member will endorse it (having each respectively rejected one premise) (Pettit, 2004, p.171).

Pettit suggests that a better way to overcome this dilemma would be for an institution to opt for the first option, i.e., premise-centred reasoning. By basing their final view on what they think about just the premises under consideration (i.e., just *p* and *q*) and inferring what a rational agent would think from there, an institution would be able to consistently say that it believed *p*, believed *q*, and believed *p and q* – even though the majority of members do not actually believe *p and q*. Again, this causes problems for a summative view of group mental states, as the institution develops its own mental states that are not necessarily reflective of those of its members.

List and Pettit think that meeting basic requirements for rationality is essential for attributing agency to institutions. They suggest that taking note of previous espoused beliefs on other issues must also be taken into consideration. For example, if the institution decided to ban the use of plastic straws because of concerns about the environment in the past, it would be inconsistent of them, and irrational, to wrap some of their goods up in single use plastic, even if on this occasion all board members think it is a good idea to do so. Thus, it becomes even more difficult to regard the views of institutions as a mere aggregation of individual members' views. If membership of an institution were to change over time, it may be the case that an institution that decided not to use plastic straws is now full of members that are all in favour of using single use plastic. Yet as the identity of the institution remains the same, the fact that its integrity is stake would impact the views of the institution going forward. Given the demands we put on institutions to be rational, their beliefs cannot be a simple aggregation of the views of the institution's members. As we have seen, such aggregation can yield inconsistent results. In this regard Pettit notes:

> The hard choice that a group in this dilemma faces is whether to let the views of the collectivity on any issue be fully responsive to the individual views of members, thereby running the risk of collective inconsistency; or whether to ensure that the views of the group are collectively rational, even where that means compromising responsiveness to the views of individual members on one or another issue. You can have individual responsiveness or collective rationality but you cannot have both – at least you cannot have both for sure. (Pettit, 2004, p174)

This also gives us reason to think that Gilbert's joint commitment might be sufficient for attributing a belief to a group like an institution, but will not be sufficient for attributions of agency. The members of a minded institution will have made a joint commitment when joining the institution (e.g., in their contracts they will have agreed to align themselves with the purpose of the institution), and they may continue to jointly commit to new institutional

beliefs – but there is also a need for institutional decision-making and belief formation to follow the norms of rationality.

Peter French also notes that institutions may well have mental states that differ from those of their members. In this regard he notes that "corporations have reasons because they have interests in doing those things that are likely to result in realization of their established corporate goals regardless of the transient self-interest of directors, managers, etc." (French, 1979, p214). He reiterates, "we can sometimes describe […] events as corporate and still further as done for corporate reasons that are qualitatively different from whatever personal reasons, if any, component members may have for doing what they do" (French, 1979, p215). Thus French too separates the individual beliefs, desires and goals of members from that of the institution – and indeed he suggests, in much the same way as Gilbert does, that the behaviour of individuals within the institution will inevitably be constrained and subordinated by the overarching views and policies which govern the institution (French, 1979, p213). Consequently members of institutions would make sure that any views they personally held which were inconsistent with the views espoused by the institution would not transfer to the institution itself. For example, if an employee is interviewed by the media and asked what their institution thinks about *p*, if the employee's personal belief about *p* is at odds with the institutions belief about *p*, they would not contradict the view of the institution and say what they personally believe in that interview.

If one returns to Tollefsen's suggestion that we can successfully adopt the intentional stance toward institutions, part of what will make it possible to see institutions as intentional agents is our ability to see them as behaving in a meaningful way over time. If an institution did decide at some point to act in accordance with their view that they believe *p,* and then at a later point in accordance with their belief that *q*, and then later still to act in accordance with the view that they do *not* believe both *p and q*, then their behaviour would start to look random – like a malfunctioning, or unthinking, robot – and we may be unable to discern a pattern in their behaviour that would make adopting something like the intentional stance towards them possible. Given that the public demands consistency from its companies/NGOs/governments etc., and calls these institutions out when there are failures in this regard (see for example Wagner et al.'s [2009] paper on corporate hypocrisy), there is some reason to suppose that we do indeed see these institutions as rational agents. What's more, institutions aren't just being called out for behaving inconsistently – they also appear to be capable of behaving rationally (i.e., adopting the intentional stance towards them works). From the arguments above we can see that institutions can be the bearers of mental states that

cannot be appropriately attributed to any individual members. As a result, we now have reason to suppose that our strategy of adopting the intentional stance towards institutions is indeed working for the right reasons.

Section 3: Institutions can be Agents

Once one is able to see institutions as capable of having mental states like beliefs and desires, one has all the building blocks one needs to see institutions as genuine agents. Indeed, actions are often identified as such due to their stemming appropriately from a belief and desire pair – a pair which according to Donald Davidson provides the 'primary reason' for an action (Davidson, 1963, p13). Christian List and Philip Pettit outline what they take to be the basic, metaphysically neutral, requirements for what it takes to be agent, and suggest that an agent needs to possess the following features:

> First feature - It has representational states that depict how things are in the environment.
> Second feature - It has motivational states that specify how it requires things to be in the environment.
> Third feature - It has the capacity to process its representational and motivational states, leading it to intervene suitably in the environment whenever that environment fails to match a motivating specification. (List & Pettit, 2011, p20)

Thus for List and Pettit, being an agent is to have representational states – something like beliefs about how the world is – for example, a philosophy department can have beliefs about how many students have enrolled in its courses. In addition, an agent will have motivational states – something like desires for things to be a certain way – for example, a philosophy department might want more students to enrol in its courses. An agent will also have a capacity to process these states and to act accordingly, changing how they perceive things to be to how they want them to be. For example, a philosophy department might put more effort into advertising the courses it offers, or change the courses it offers in order to have wider appeal. List and Pettit do add that the idea of an agent will also need to incorporate some kind of 'standard of rationality' (List & Pettit, 2011, p24). We saw in the previous chapter that part of what one does when one adopts the intentional stance is to assume rationality, and in the previous section I argued that we expect some degree of rationality from institutions. For List and Pettit, in order for us to say of a being that what it did was an action at all, one would need to see that what it did was, for example, consistent with its other actions, in order for us to be able to rule out the possibility that its behaviour was just random.

List and Pettit happily endorse a functionalist view regarding mental states, saying that when it comes to ascribing agency to some or other entity, they are merely concerned with whether or not that entity has the relevant capacities that play the appropriate roles:

> We make no assumptions about the precise physical nature of intentional states. They may be of a wide variety of kinds. They may be electronic or neural configurations of the agent, for example, depending on its robotic or animal nature. They may be localized in the agent's brain or central processing system or dispersed throughout its body. We only require that they be configurations of the agent – or perhaps configurations of the agent in context – that play the appropriate functional role. (List & Pettit, 2011, p21)

Thus List and Pettit are neutral about the kind of thing that could instantiate the relevant states needed to be an agent. Indeed, they open their discussion on the requirements of agency by outlining an example of a little robot that picks up cylinders and places them in an upright position, remaining inactive if all the cylinders are upright, and going to a cylinder and propping it up again if it falls over (List & Pettit, 2011, p19-21). For List and Pettit, as long as there are states that play the relevant roles of belief, desire and intention, then the being in question counts as an agent, even if only a simple one (if the robot is capable of observing the environment it is in and forming a representation of that, if it is capable of representing how it wants things to be, and if it is capable of acting in such a way that it can bring about changes to that environment in accordance with how it wants things to be, that is, roughly speaking, good enough for agency).[37]

Pettit suggests that we can get a good sense of whether or not some or other system is an agent by observing it (Pettit, 2014, p15). For example, in the case of the robot if we see that over time it behaves in a way that seems to indicate that it wants cylinders to be upright (i.e., it picks up any that fall down, it doesn't pick up things that aren't cylinders etc.), then we can infer that it has certain, fairly reliable, representations of how the world is – in other words it can tell, fairly reliably, barring conditions of pitch darkness, for example, what cylinders are, and whether or not they are in an upright position. Moreover, because it lifts up cylinders that fall over, and leaves cylinders that are standing upright alone, we can infer that it has some aim (namely to have all the cylinders in its surroundings be in an upright position). From this we can infer that the robot is a kind of simple agent, that it has representations of the world, desires the world to be a certain way, and acts to bring that

[37] Although some have suggested that these notions are in particular personal, not merely agential (see, for example, Davidson in 'Thought and Talk' [1975]). This suggestion is discussed further later in the chapter.

about. Indeed, this conclusion is one that Tollefsen reaches with regard to adopting the intentional stance toward institutions.

Pettit does acknowledge that this isn't an infallible way of determining whether or not some or other system is an agent, as the robot could, for example, unbeknownst to an observer, be remote-controlled by someone else who constantly 'tells' the robot what to do, or the robot may have been so carefully pre-programmed for every possible eventuality that it was likely to encounter, that it may just seem to be exercising agency (Pettit, 2014, p5). Thus he adds further that to count as a genuine agent, the goals that the system has must come from the system itself: "if the system is enduringly organized within itself so that it displays the required pattern of behavior, then there can be no room for doubt about its agential status" (Pettit, 2014, p6). One may be concerned that when it comes to institutions, the goals they have come from individuals that make up the institution rather than the institution itself. Yet as we saw above, institutions are capable of having beliefs of their own, beliefs that in some cases no individual member of the institution has. Institutions are in a sense dependent on their members – "whatever representations are supposed to guide the group in its fidelity to those patterns, they have to be formed on the basis of representations formed in some or all of its members: the members, after all, are the group's eyes and ears" (Pettit, 2014, p26). In spite of this, that an institution acts in a unified and coherent manner over time will be due to its organisational and decision-making structures, and not something inherent in individual members (which are fungible, after all). For Pettit then it is clear that institutions can meet the same minimal requirements for agency that the robot can.

Echoing again the concern that we may tend toward chauvinism in refusing to ascribe mental states or other features we ordinarily reserve for the exclusive domain of human beings to others, Harry Frankfurt notes:

> [W]e must be careful that the ways in which we construe agency and define its nature do not conceal a parochial bias, which causes us to neglect the extent to which the concept of human action is no more than a special case of another concept whose range is much wider. (Frankfurt, 1978, p162)

Perhaps it should be noted that List and Pettit are not the only ones who insist that the physical make-up of an agent shouldn't matter when ascribing intentional attitudes to it. Again, we may well be displaying the kind of chauvinism Ravenscroft warned of in assuming that only human beings, or creatures sufficiently similar to human beings, could manifest agency. Isaac Levi and Dennett and Carol Rovane make similar points:

> Perhaps, differences in the 'hardware' should make a difference in the view we take of principles of rational preference, belief, valuation and choice. But unless a decisive case is advanced that this should be so, it seems sensible to seek an account of rational choice, belief, preference and valuation which is indifferent as to whether the agent is human, automaton, animal, angelic or social. (Levi, 1990, p152).

> The concept of an Intentional system is a relatively uncluttered and unmetaphysical notion, abstracted as it is from questions of the composition, constitution, consciousness, morality, or divinity of the entities falling under it. (Dennett, 1971, p100)

> Human size is not basic to agency. (Rovane, 2004, p181).

Nevertheless, many philosophers remain very reluctant to ascribe agency – anything like the kind of agency we think of human beings as possessing – to List and Pettit's robot, and indeed to other entities that appear to be agent-like (such as some non-human animals or institutions). Indeed, taking note of the seemingly meaningful behaviour of many non-human animals, Davidson suggests that they are capable of behaving appropriately in their environments without actually possessing what one could rightly call thoughts, insisting that:

> The fly serves to remind us that an organism can discriminate aspects of its environment with superb accuracy, and make use of the resulting information in complex ways that help keep it alive, without anything we would, or should, call thought. (Davidson, 2001, p135)

Davidson goes on to note that, given this seemingly intelligent behaviour:

> it is important to reflect on why none of this shows they have propositional attitudes: beliefs, desires, doubts, intentions, and the rest. Dumb beasts see and hear and smell all sorts of things, but they do not perceive that anything is the case. Some non-human animals can learn a great deal, but they do not learn that something is true. (Davidson, 2001, p136)

Davidson sees thought as something which requires language, as he argues that without language one will not have the necessary conceptual framework for propositional attitudes (Davidson, 2001, p137). And propositional attitudes, like beliefs and desires, are necessary for genuine agency as opposed to mere behaviour. Thus, for Davidson, List and Pettit's robot (along with flies and other 'dumb beasts') may appear to behave in a meaningful way, however, lacking language it would fail to count as a genuine agent. Perhaps something resembling this concern can be extended to the claims that groups like institutions could be agents. Indeed, even though institutions may appear to have "agent-like features" (Pettit,

2014, p2), there is certainly some suspicion that this is simply just an appearance.[38] Such a concern does not seem to hold, however, when one considers the number of utterances (written and verbal) that we attribute to institutions rather than individual members of those institutions. Indeed, Miranda Fricker draws attention to the ubiquity of communication on the part of institutions, noting that "[w]e gain information from collective, often institutional bodies all the time—from the publications of committees, news teams, or research groups, from web sites such as Wikipedia, and so on" (Fricker, 2012, p249).

Speaking directly to Davidson's concern regarding what is necessary for genuine agency, Pettit suggests that there are "two modes of agency" (Pettit, 2014, p9): what he describes as non-personal agency, and personal agency. He views his example of the simple robot, as well as non-human animals, as having non-personal agency (i.e., they can have representations and act on those representations to meet certain goals etc.), while arguing that human beings are capable of personal agency – they possess the ability to do a number of things not available to the robot or a dog, say, such as the ability to reflect on and review beliefs and attitudes. In other words, they have conceptual frameworks and are capable of the kind of thought Davidson sees as essential to agency.

One distinctive feature of a personal agent that Pettit picks out is that such agents are 'conversable'. In other words, they are capable of verbalising their thoughts, which ties into Davidson's concern with the connection between language and thought. Again, while Pettit clearly allows for the possibility of non-personal agents which lack the ability to converse (and even lack language), he notes that in order to see an entity as a personal agent the capacity to converse is important. Indeed, he notes that "being a personal agent, you will be expected to prove yourself a conversable agent too: someone we find it possible to reach in the realm of words and to engage to our mutual benefit." (Pettit, 2014, p19-20) – and he seems to view this as a pretty good test of whether or not someone or something will be regarded as a personal agent. Moreover, while he has noted that we can infer agency in simple robots or animals from impersonal observation, he suggests that in the case of personal agents, like human beings, we can also gather evidence of agency via interpersonal interaction – even suggesting that if we are utterly unable to communicate with someone, then we will tend to doubt whether or not they are in fact bearers of personal agency. From

[38] Recall, for example, Quinton's (1975) suggestion that all talk of groups having mental states or performing actions is metaphorical shorthand for attributing these mental states and actions to a number of individuals. See also Kirk Ludwig's (2016; 2017) arguments for an account of the agency of institutions in terms of individuals, noting that while there is a "superficial appearance that there is a group agent […] only its members are agents of anything" (2016, viii).

here Pettit suggests that institutions are personal agents – agents who are not only capable of forming representations about the world, and who have goals they desire that they work to achieve, but who's agency we can explicitly detect because they are conversable: institutions are language users, they tell us what they believe and desire etc., they interact with us. Often they do so via their spokespersons or via decision-making boards, but (at least in the right context) the views expressed are attributable to the institutions, not the individuals acting as the mouthpiece of the institution. Thus institutions seem capable of meeting even the more stringent requirements for agency set out by Davidson.

Section 4: Institutions can be Moral Agents

Clearly it is possible to draw the distinction that Pettit has between the agency his robot, or a non-human animal, might have and the kind of agency that we think of ordinary human beings as possessing – i.e., the kind of agency that goes hand in hand with moral responsibility. I have claimed Pettit is correct to think that institutions, like individual human beings, are capable of personal agency. Yet personhood is often cited as a distinguishing feature when it comes to incorporating some, and not others, into a moral community. Although institutions are granted the legal status of persons in many parts of the world (see The Yale Law Journal Company, Inc., "Constitutional Rights of the Corporate Person", 1982), French has elaborated on the distinctions between metaphysical, moral and legal personhood (1979). He argues that it is thus possible that one need not ascribe metaphysical (or moral) personhood to institutions despite their legal status as persons. Nevertheless, some philosophers like Carol Rovane have argued that "the concept [person] is to be applied on normative grounds, without metaphysical prejudice" (Rovane, 1994, p393), and thus the idea of moral personhood should be a concept that is not restricted to being the exclusive endowment of human beings. Indeed, she strongly emphasises the possibility that groups who meet certain requirements could count as moral persons too.[39] In a similar vein, Adam Kadlac has argued that the concept of a person picks out certain morally relevant properties "such as rationality and self-consciousness" (Kadlac, 2010, p422) that need not in principle be confined to human beings, while noting that exactly what properties we think are constitutive of personhood have been the matter of some debate. Thus it would seem that in order for a

[39] While the view of personhood that Rovane puts forward is in opposition to views like that of Eric Olson – namely the "Biological Approach" [1995, p160] – Olson's claim that we are animals does not rule out the possibility of there being other non-human persons (Shoemaker, 1999, p496).

non-human like an institution to count as a person one would need to show that they possessed whatever we think of as "the relevant person-making properties" (Kadlac, 2010, p422). Consequently, in spite of the fact that some countries do recognise certain institutions as persons, more needs to be said to determine whether or not institutions could, or should, count as persons. In what follows I will attempt to argue that institutions do possess properties widely considered to be constitutive of personhood – however, more importantly, I argue that even if we do not wish to grant them personhood per se, we can and should still see them as moral agents.

Dennett notes "whatever else a person might be – embodied mind or soul, self-conscious moral agent, 'emergent' form of intelligence – he is an Intentional system, and whatever follows just from being an Intentional system thus is true of a person" (Dennett, 1971, p100). So being an intentional agent, being able to be interpreted as an agent, is at least part of what a person is. This is so because the rationality that one assumes when adopting the intentional stance seems to be important for seeing an agent as acting morally too. Yet given that there are plenty of intentional systems that we would not classify as persons (arguably supercomputers and trees, and dogs and cats, and perhaps institutions), what else are persons such that they count as members of the moral community while these other intentional systems do not?

One suggestion that has already been touched on is the fact that persons are capable of propositional attitudes – they not only have beliefs and desires etc., they have the capacity to reflect on those. Indeed, Christine Korsgaard regards this as the hallmark of morality. She begins by noting that, "[a] lower animal's attention is fixed on the world. Its perceptions are its beliefs and its desires are its will. It is engaged in conscious activities, but it is not conscious of them." (Korsgaard, 1996 p78). However, she argues that human beings are capable of more:

> [W]e human animals turn our attention on to our perceptions and desires themselves, and we are conscious of them. That is why we can think about them […] our capacity to turn our attention onto our own mental activities is also a capacity to distance ourselves from them and to call them into question. (Korsgaard, 1996, p78-9)

And it is this capacity for self-reflection, the capacity to be aware of what we believe etc., and to question those beliefs, that propels us into the moral sphere.[40] Yet given that institutions

[40] See also Frankfurt, who suggests that "People are commonly concerned about their motives; they want their actions to be motivated in certain ways, and not in others. Insofar as they find certain of their own motivational

are also capable of propositional attitudes, we may well be justified in viewing them as the appropriate subjects for the attribution of praise and blame. In other words, because they can assess and review the beliefs or attitudes that they hold (the philosophy department that would not accept late assignments for any reason can reassess whether or not that is a fair and reasonable policy, for example), we can see them as responsible for that belief or attitude.

While it may not be necessary to see institutions as persons in order to see them as moral agents, the work of Ifeanyi Menkiti on the conception of personhood and its essential link to morality support the view that we ought to see institutions as bearers of moral responsibility. With regard to African traditional societies, Menkiti writes, "[John Mbiti] notes that an individual's relation to the group is best described as one of "I am because we are" (1970: 141)." (Menkiti, 2004, p324). In other words, reciprocal relationships are important for the development of personhood – someone becomes a person through other people, through being part of a community, through being treated as a person, and fulfilling a person's roles (i.e., moral roles). In this vein, Menkiti suggests that "morality demands a point of view best described as one of beingness-with-others" (Menkiti, 2004, p324). There are parallels to be made here with Dennett's personal stance (1988, p178). Like Dennett, Menkiti is asking us to look at our practices, and how that is related to or generates incorporation into society, and one's moral agency – what one might take to be constitutive of personhood.

Institutions also exist within a social context, where certain demands are made of them, and thus they could, in principle, also embody 'beingness-with-others'. As List and Pettit note, evidence for this is the fact that institutions frequently enter into contracts with one another, and with individual members of the public, and as such incur obligations in that regard (List and Pettit, 2011, p176). Moreover we can see further evidence of their incorporation into moral communities reflected in the increasing demands for certain institutions to behave ethically. For example, businesses are now being called upon to stop selling products in single-use plastics, and many are bowing to pressure and are heading the call (e.g., according to an article from BBC News, "More than 40 companies have signed up to a pact to cut plastic pollution over the next seven years. The firms, which include Coca-

tendencies objectionable, they attempt to weaken and to resist them. They accept and identify with only some of the desires and dispositions that they encounter in themselves" (2004, p19). When we are moved to act by desires that we do not endorse then we can be said in a sense to act against our will as we do not will that we are so moved to act. On the other hand, when we act on desires that we are upon reflection happy to be moved by, then we have acted freely. For Frankfurt then, self-reflection also gives us our freedom (or at least "as much freedom as it is possible for us to conceive" (2004, p20). It seems that this will be a feature of those we wish to hold morally responsible too.

Cola and Asda, have promised to honour a number of pledges such as eliminating single-use packaging through better design." [Shukman, 2018]). Recently Facebook has faced criticism over privacy violation issues, and a failure to control the propagation of 'fake news' on the platform (e.g., according to Mark Schaefer, "Over the past five years, they have violated so many ethical standards that it's hard to even keep track of it […] it seems like Facebook has no moral compass" [2018]). In additional news articles, there has also been a sharp focus on corporate responsibility for climate change, which has been highlighted by global protests (Fraser, 2019). There is an acknowledgement that various institutions, including Coca-cola and Facebook, have moral duties towards us, and there exists the expectation that they are capable of making (and keeping) promises, and that they ought to have a moral compass. Menkiti reiterates that "in the stated journey of the individual toward personhood, let it therefore be noted that the community plays a vital role both as catalyst and as prescriber of norms" (Menkiti, 2004, p326), and again, we do think of institutions, in addition to human individuals, as being located within communities that prescribe norms, and we may well recognise institutions as moral agents by treating them as such. Indeed, Dennett also notes that "whether something counts as a person depends in some way on an attitude taken toward it, a stance adopted with respect to it" (Dennett, 1978, p177). In other words, treating human beings as persons is partly constitutive of their personhood.

Menkiti draws a similar distinction to that of Pettit between a non-personal agent and a personal agent, or a mere agent as opposed to a moral agent, in the following observation:

> [C]onsidered as an individuated source of consumption, a bundle, as it were, of primary appetites, the individual could still count as an agent in the world. But to go beyond the raw appetitive level to the special level marked by the dignity of the person, something more would seem to be needed. In this regard, 'individual' and 'individual person' may carry somewhat different weight and it is the context of discussion that spells out whether they converge or diverge. (Menkiti, 2004, p325)

This should perhaps be seen in conjunction with the fact that Menkiti suggests that personhood is also something that a human being could fail to achieve (p326) – and thus that *actually* behaving ethically, as a person should, is what grants one personhood.[41] Yet this observation could serve to emphasise again that not all groups are necessarily moral agents – that certain capacities are required. However, institutions – purposive, possessors of mental

[41] Although it is possible that, in addition to one failing to achieve the status of personhood because one fails to meet the requirements in some respect, it is also possible that this may occur because the community fails to recognise one as a person.

states like beliefs and desires, capable of reflection and conversing and acting in the world – would appear to occupy a space in the moral sphere. They are capable of behaving both ethically and unethically, and are the appropriate targets of praise and blame. Indeed, French explicitly argues for seeing corporations as moral persons, "as members of the moral community, of equal standing with the traditionally acknowledged residents: biological human beings" (French, 1979, p207) precisely because they can be seen as agents, acting for reasons of their own, and which, as a result, can rightly be held to account for their actions (French, 1979, p215). Nevertheless, Menkiti does also speak of the fact that, as the traditional African view of personhood is conceived of as a progression, that personhood is something that is attained, "the older an individual gets the more of a person he becomes" (Menkiti, 1984, p173). Thus although Menkiti talks of the "moral arrival" (2004, p326) of a young adult, he also talks of younger persons as exhibiting "a tendency towards self-centeredness in action, a tendency to see the world exclusively through their own vantage point" (1984, p175), marking them as possessing a lesser degree of "excellencies considered to be definitive of full personhood" (1984, p173) than elders in a community. We may well be able to view institutions in a similar light to younger adults here – i.e., less morally astute, and thus not full persons, but still participants in the moral sphere. Indeed, noting that certain institutions have fallen so short when it comes to acting morally, there has been some suggestion that they be regarded as psychopaths. For example, Ted Steinberg argues that "the largest corporations have market capitalizations that are bigger than the gross domestic products of some nations. That economic power […] has been used to cause untold social dislocation and ecological turmoil, just as you would expect of a psychopath" (Steinberg, 2007, p372). In less extreme scenarios we might consider the fact that the notion of moral maturation could apply just as well to institutions as it does to individuals. When a new business is set up, it will not necessarily be trusted by its community – whether by customers who do not know what level of service or products they will receive should they frequent that business, or by suppliers, who may insist on a new business paying upfront for goods for a certain amount of time before it would qualify for credit, etc. Institutions could then reach a higher degree of moral maturity via reliable and honest behaviour over time. [42]

[42] On Menkiti's view, granting personhood to an individual (or an institution) will depend upon the norms of the society and the recognition of the society. This may be problematic. When the status of personhood is something that is conferred by one's society, there is always the risk that it can be withheld or taken away unjustly. For example, one may feel that a certain religious institution ought to be recognised as a legitimate participant in a community, and yet if that community does not recognise them, they will not be regarded as such. I would suggest that one can accept the validity of this concern, and allow for the possibility that

Institutions are certainly often capable of having significant effects upon the social world – and things can go wrong (the unethical behaviour of Facebook cited above is just one example). Yet in institutional settings it is not always obvious who is ultimately to blame for unethical practices. In many legal cases involving corporations, courts have concluded that no individuals within the company can be held truly accountable, and so no one at all is held responsible – for example, the notorious case of the Herald of Free Enterprise ferry (Pettit 2007). Due to negligence by the crew, the ferry sank in 1982, leading to the deaths of 193 people. However, as no individual had done anything serious enough by themselves to contribute to the sinking of the ferry, no one at all was adequately penalized for this preventable tragedy. Examining the possibility of institutions being seen as moral agents is of particular importance if, as was the case with the Herald Free Enterprise, we risk a dearth in responsibility if institutions are not held accountable for their actions. Seeing institutions as moral agents allows us to make sense of blaming the institution even when no member of the institution is to blame. If we can attribute a failure or wrongdoing to an institution, even when no individual within that institution is guilty of any failing or wrongdoing, this could help us address situations more fruitfully where a failing or wrongdoing still took place.

As a final note, Pettit also argues explicitly for institutions to be regarded as persons (in the moral as well as legal sense), given that they are capable of meeting what a number of theorists suggest are the relevant criteria for moral personhood. He does so while acknowledging nevertheless that such institutional persons will have limitations we do not ordinarily consider as features of natural, or human, persons:

> Institutional persons form their collective minds only on a restricted range of matters, to do with whatever purpose they are organized to advance. […] Integrated collectivities are persons in virtue of being conversable and responsible centers of judgement, intention and action. But they are persons of a bloodless, bounded, and crudely robotic variety. (Pettit, 2004, p188)

Such an acknowledgement may go some way to explain the instinctive recoiling one may feel on considering such entities persons, and thus taking Menkiti's view of personhood as acquired, and as existing to varying degrees, may be helpful here. In any case, it would seem

---

sometimes the community will get it wrong. To correct this the procedures by which a community either recognises or fails to recognise as persons some or other individual or institution ought to come into focus, and whatever the criteria are by which a community judges individuals or institutions as lacking should be interrogated. Any criteria based on certain kinds of features such as skin colour, or political affiliation, or religion, for example, should not be included.

we have good reasons to think that not only are institutions genuinely capable of being minded agents, they are capable of being moral agents too.

Chapter 4: Institutions can be Liars

"BP lied about size of U.S. Gulf oil spill, lawyers tell trial" – Kathy Finn, *Reuters* (2013)

So far I have established how it is we can view institutions as morally-culpable, minded agents, and gestured towards why it can be useful to do so. The remainder of my thesis will illustrate exactly why it is useful to regard institutions as such through discussing a number of situations in which institutions can be said to behave badly, sometimes with disastrous consequences. Once we see institutions as legitimate members of the moral community, claims that institutions lie, for example, make sense – and we should take such claims seriously. Throughout recent decades a number of institutions have been accused of acting unethically, and often of lying or deliberately misleading the public. For example, in 1978 a tragic motor accident occurred on a highway in America – a van crashed into the rear end of a Ford Pinto, causing the gas tank to rupture, and the Ford Pinto to burst into flames. Three young women who were in the car died in that fire. Following this incident Ford was charged with three counts of reckless homicide, the first ever such charge to be brought against an American corporation (Hoffman et al., 2014, p139). The suggestion was that Ford "knowingly and recklessly" went ahead with what it knew was an unsafe gas tank design and placement in the Pinto to increase profits (Hoffman et al., 2014, p139). More recently in 2010 BP is said to have lied about the amount of oil spilled in the Gulf of Mexico (Finn, 2013), and even more recently in 2013 scandal broke in a number of European countries following the discovery in Ireland "of horse meat in frozen burgers labeled 'pure beef'" (Jarry, 2019). Firms such as Spanghero in France have since been brought to trial, accused of deliberately trying to pass horse meat off as beef, with investigators alleging that "Spanghero knew it was buying frozen horse meat and switched the customs code on the packaging" (Jarry, 2019), an accusation that Spanghero had denied (Vinocur, 2013).

As with other group level phenomena, the exact way in which we ought to understand institutions lying has come under some scrutiny. However, before looking more carefully at how institutions lie, I would like to begin this chapter by briefly looking at lying in general. In Section 1 I start with a discussion of the norms associated with making assertions, namely a commitment to truthfulness, and how we can understand lying as a violation of these norms. Once we have a grip on the key aspects of lying, we will be in a better position to see how institutions, as well as individuals, are capable of lying in Section 2. I look in particular at Jennifer Lackey's discussion of group lies in a recent paper, and her own account of group

lies. While I am not entirely satisfied with Lackey's account, it is notable insofar as she is primarily concerned with drawing a clear distinction between the lies of a group and the lies of individuals.

Yet it is not just the case that lying can be seen as a violation of the norms of assertion. One primary reason we have not to lie stems from a concern for those who are deceived, that those who are lied to are actually harmed in some way. Thomas Hill suggests that lying inappropriately influences the behaviour of others by providing false information, or selective information, or by withholding information etc. (Hill, 1991, p31),[43] and Bernard Williams suggests that we can see lying to others as a form of manipulating or using them (Williams, 2002, p93).

While institutional lies would also seem to similarly violate a kind of tacit commitment to communicate honestly, the harms associated with institutional lying are the same as those associated with individual lies, if not more severe insofar as institutional lies are likely to reach, and so affect, far more people. The behaviour of those lied to is often altered due to the false, or selective information put forward by institutions, or by their withholding information – for example, when companies switch labels on products in an attempt to pass horse meat off as beef, it can affect consumers all over the world. Similarly, if a pharmaceutical company lies about the dangers associated with a new drug, vast quantities of people could be extremely adversely affected. Institutional lying can also do further indirect harm insofar as the trust of the public is lost when they discover that they've been lied to. Indeed, Sisella Bok highlights the corrosive effects of lying rather well when she notes:

> trust is a social good to be protected just as much as the air we breathe or the water we drink. When it is damaged, the community as a whole suffers; and when it is destroyed, societies falter and collapse. (Sissela Bok, 1978, p26-27)

A further reason not to lie is that doing so often goes against one's own self-interest. Insofar as lying damages one's reputation it can be in one's own interest to be honest, as it may be difficult to get others to cooperate with one if one has a reputation for being untrustworthy (Williams, 2002, p94).[44] Moreover, while there are doubtless advantages to

[43] Hill even suggests that even benevolent lies, lies which are told supposedly for the benefit of those lied to, are bad, as they involve not taking the autonomy of the person lied to into consideration (1991, p25).
[44] Interestingly, Brad Sagarin, Kelton Rhoads and Robert Cialdini suggest that there is another important reason not to lie – one which involves concern for the liar as well as the one being lied to. They have carried out tests which show that even in cases in which a lie is not discovered, "a liar will begin to perceive the recipient of the

lying (such as avoiding the negative consequences of one's actions), the fact that lying can also go against one's own self-interest applies equally well in the case of institutions as in the case of individuals. Given that so many institutions (if not all), be they corporations or universities, hospitals or the government etc., have to interact with the public, or at least with other institutions, having a good or a bad reputation will affect those relations. If an institution becomes known for lying, they may well lose business (a number of food outlets were undoubtedly sceptical of purchasing beef from Spanghero, for example) or power (as when a government becomes associated with lies and corruption). Knowing that one can rely on institutions to be honest, that one can trust them, is important for societal well-being, and important for said institutions. Thus, as Michael Aßländer points out, being honest can actually be good for business (Aßländer, 2013, p758).

Given the harms associated both with being lied to and with lying, one may well wonder about the harms of 'lying to oneself' or self-deception. These will be topics I turn to in forthcoming chapters. I discuss how we can make sense of institutional self-deception and an institution lying to itself in Chapter 5, before moving on to a discussion of the harms and moral implications of this, and some potential solutions in Chapter 6.

Section 1: Norms of Assertion and Lying

It is widely acknowledged that, as Roderick Chisholm and Thomas Feehan write, "the concept of assertion is essentially *normative*" (1977, p151). A number of suggestions regarding the nature of these norms have been put forward. For example, Chisholm and Feehan claim that a hearer is justified in taking the asserter to believe what they assert (1977, p151-2), which implies we think that one ought only to assert what one believes. Others have put forward views that an assertion is only proper if it is true (Weiner, 2005). In her paper, 'The Norms of Assertion', Jennifer Lackey notes that a common view in this regard is that one ought only to assert what one knows (i.e., knows to be true) (Lackey, 2007, p594). While Lackey herself does not endorse such a stringent norm of assertion, she does propose a

lie as less honest as a result of the liar's own deception" (Brad Sagarin et al., 1998, p1167). They suggest that this could have three different causes, namely a) a false consensus mechanism, whereby one falsely believes that others are like you – in other words, liars, b) an ego-protective mechanism, whereby one comes to believe that others would behave in the same way as oneself – in other words lie in similar circumstances, or c) a belief in a just world, whereby one comes to believe that the person who has been lied to must have deserved it in some way because bad things (here the harms associated with being lied to) don't happen to good people (1998, p1167). Notably their research suggests that if one tells lies to others, one comes to see those others as deserving of mistrust, and even of dislike and scorn (1998, p1167). Thus, insofar as we think these inclinations and views of others negative ones, telling lies would seem to harm the liar.

"reasonable to believe norm of assertion" (Lackey, 2007, p608), suggesting that an assertion is only proper if one asserts what it is reasonable to believe is the case. In defense of this less stringent norm of assertion, Lackey suggests that such a norm can account for instances where one may not, strictly speaking, know what one asserts, but where such assertions are still deemed appropriate – thus she notes "there is a clear sense in which speakers who assert reasonably believed falsehoods and Gettierized beliefs are not subject to criticism" (Lackey, 2007, p597).

The 'reasonable to believe' norm of assertion that Lackey endorses is also meant to account for the possibility of "selfless assertion" (Lackey, 2007, p598), whereby an asserter is able to make a proper assertion – one that is free from criticism – even if the asserter does not personally believe or know what they assert. Lackey cites examples of persons who, perhaps because of acknowledged bias or prejudice, just cannot or will not believe in accordance with the evidence. However, recognising that this is so, they assert what does in fact accord with the evidence rather than what they personally believe to the case – for example, when a racist juror "can't shake the feeling that the man on trial is guilty" (Lackey, 2007, p598) even though the evidence does not support this conclusion. If the juror goes on to assert that the man isn't guilty (despite personally thinking they are because of racial prejudice), this assertion should be regarded as proper according to Lackey (who notes that such a person may well be criticised for their own actual belief, but not for their assertion, which is deemed appropriate). Thus, Lackey concludes that "it is a mistake to require proper assertion to pass through the doxastic states of the asserter" (Lackey, 2007, p600).

Bernard Williams' writes that, from an enquirer's perspective, when "genuinely asking a question, wondering how things stand, I aim at a true answer" (2002, p84). This claim is similar to Lackey's. She is also attempting to put forward a norm that allows for situations in which one's primary aim is to provide a 'true answer', even though one can personally fail to believe, or know, what one asserts. Thus, while Lackey may appear to weaken the norms associated with assertion, she does so in order to accommodate situations in which one appropriately asserts something that unluckily turns out not to be true, and also to accommodate situations in which one may hold a view that one recognises most reasonable people do not (Lackey, 2007, p598-9). In such situations she thinks it is better to assert what it is reasonable to believe, rather than what one personally thinks about some matter because of acknowledged biases etc. Nevertheless, she introduces her less stringent norm of assertion in a direct attempt to preserve the link between making appropriate assertions and (something

like) imparting knowledge. I therefore take it that when making assertions there tends to be an expectation on the part of hearers that one will be truthful, or provide a 'true answer'.

When it comes to lying, Timothy Williamson suggests that "someone who knowingly asserts a falsehood has thereby broken a rule of assertion" (Williamson, 1996, p489). I take it that Lackey's juror who states what they do not personally believe to be the case is nevertheless attempting to avoid asserting a falsehood. Thus the expectation we have regarding assertions – that the asserter is honest, that they will provide a truthful answer – is violated when someone lies. Indeed, Thomas Hill argues that "lies are often wrong at least in part because they are breaches of a promise to be truthful" (Hill, 1991, p27). While Hill claims that even though we may not make *explicit* promises to one another to be truthful, as noted above, there exists an expectation that one will endeavour to make honest assertions – to say what one believes is the case, or at the very least, what one believes is broadly regarded as being the case (and so likely to be true) when one is aware that one's own personal view on some matter is genuinely unreliable – and thus lying violates the norms we think are associated with making assertions.

Some theorists have suggested that lies explicitly involve asserting falsehoods, and that lying is thus just one form of deception among many (Chisholm & Feehan, 1977; Bok, 1978).[45] Others have argued that lying need not only occur when making assertions. Sam Harris, for example, argues that "to lie is to intentionally mislead others when they expect honest communication" (Harris, 2011, p4) – such a definition would incorporate, for example, lies of omission. Deliberately failing to say something when one knows that someone will form a false belief as a result is thus regarded as a kind of lie. The implications of what is said can also play an important role in deception, and thus one can still be dishonest even if what one explicitly says is, technically speaking, true. As Williams notes, "all sorts of inferences may be reasonably made from what people say and the ways in which they say it" (2002, p99), and the speaker can deliberately exploit these expectations without telling an outright lie.[46] Thus it is widely accepted that such situations are also morally suspect. A similar sentiment was expressed by Damian McBride, a former special adviser to

[45] Chisholm and Feehan see lies as involving necessarily involving assertion with the intention to deceive – although they acknowledge these assertions needn't be spoken, but could include written assertions, or assertions made via gestures or signals etc., while Sissela Bok similarly sees lying as "any intentionally deceptive message which is *stated*" (p14), suggesting that, given the myriad other ways in which one can intentionally mislead someone, "deception, then, is the larger category, and lying forms part of it" (p14).

[46] For example, imagine I spilled something and used a t-shirt of yours to wipe it up. After ruining it, I throw away. You then ask what happened to your t-shirt a few weeks later, and I reply that I haven't seen it in ages. While that may technically be true, it's clear that you will probably misunderstand me, thinking I mean that I don't know what happened to it, when in fact I do.

British Prime Minister Gordon Brown. He speaks of politicians "[making] an art form of lying without lying" (Philosopher's Arms Podcast, 'Lying and Misleading', 08:10).

I am sympathetic to the view that there is not a sufficiently morally significant difference between telling an outright lie, and lying by omission or by exploiting a hearer's expectations, and so am happy to endorse Harris' broader definition that 'to lie is to intentionally mislead others when they expect honest communication', which can incorporate all of the above, while acknowledging that some may be unhappy with such a broad definition. A broad definition of lying is by no means necessary, however, to separate institutional lies from the lies of their members.

Section 2: Institutional Assertions and Lies

In Chapter 3 we saw that institutions are language users, and that we frequently interact with institutions and get information from institutional assertions all the time (Fricker, 2012, p249). Such assertions can be written down in newspapers or on websites, or made via a spokesperson on behalf of the group in interviews or broadcasts etc. Fricker, who is interested in groups counting as testifiers, makes a comparable point to that made by Hill above regarding the norms that are associated with assertion, noting that "testifying is analogous to promising" (Fricker, 2012, p258). Fricker claims that part of what is entailed when a testifier asserts that *p* is that they become committed to being truthful or honest about what they say with regard to *p* (Fricker, 2012, p271). Just as we regard institutions which enter into contracts with one another, and with individual members of the public, as incurring certain obligations, we should see institutions as being bound by the same norms of assertion that we take individuals to be.

Yet given the vast array of institutions and their many purposes, it could be argued that while we apply strict norms of assertion to certain institutions (like governments), we do not in fact presume that institutions as a whole are honest. Lots of institutions may advertise their goods and services, for example, and while we are certainly outraged at blatant false advertising, I suspect there exists an expectation on many of our parts that a lot of these corporations will be liberal with what is true regarding many of the claims that they make when advertising and trying to sell their goods and services – indeed, Sissela Bok notes that although "misleading advertising" is pervasive, it is often tolerated or dismissed on the "grounds of harmlessness and triviality used for white lies in general" (Bok, 1978, p61). Yet one can still regard institutions as bound by norms of assertion. Even if we are not surprised

by dishonest advertising, we would likely welcome honest advertising, and we do expect honesty in other assertions. When institutions violate these norms (perhaps even when advertising), we simply should not regard them as honest and reliable, in much the same way that we would not regard a pathological liar as honest or reliable. Regardless, the norms of assertion still exist, and institutions are capable of meeting the norms of assertion.

With regard to the ability of institutions to make the kinds of promises to be truthful when asserting, Fricker draws on Gilbert's account of joint commitment. She suggests that we can understand group testifiers as forming a "joint commitment to trustworthiness as to whether *p*" (Fricker, 2012, p272). Once the members of an institution have jointly committed to trustworthiness as to whether *p* they can rightly be seen as a group testifier. They can be jointly committed to trustworthiness in a number of ways – they could do so explicitly in a board meeting, say, or perhaps implicitly in virtue of acquiescing to be part of the institution and agreeing to abide by certain codes of conduct or stipulations in their contracts. Fricker herself suggests that members of institutions need not always explicitly jointly commit to being trustworthy with regard to X or Y or Z – and we have seen already that Gilbert makes the same point with regard to explicit joint commitments and background, or implicit, joint commitments (1987, p200). Thus Fricker argues that many members of various institutions "jointly commit to trustworthiness, *at least tacitly* (it comes with the job) regarding what they will tell the public" (Fricker, 2012, p272 [my emphasis]). Indeed, Fricker points out that in general it would seem that a vast number of groups or institutions do so commit to trustworthiness, whether or not they are then successful in proving trustworthy:

> I suspect that any collective body whose role is likely to involve telling people things would need to have this as one of its constitutive commitments, because precisely that sort of epistemic trustworthiness would be integral to the group's role. (Fricker, 2012, p272) [47]

Even if members of institutions did not form a joint commitment to trustworthiness, insofar as institutions offer assertions at all they are governed by similar norms of assertion to individuals. However, as noted, just like individuals, institutions can fall short of, or even violate, commitments to trustworthiness. As such institutions can prove to be bad informants,

[47] There are a few possible exceptions here – while there may be certain times where some would expect honesty from institutions like MI6, there are also certain functions of the institution that run counter to honestly communicating with the public.

or even liars, by making unreliable or deliberately false assertions with the intention to mislead.[48]

In a recent paper Jennifer Lackey explicitly discusses the possibility of a group lying.[49] Given the variety of ways in which institutions can be organised, there are a number of ways that institutions could make their assertions – which includes their lies. In her paper she assesses the possibility that the lies of a group like an institution could be understood a) in terms of the lies of individual group members, or b) in terms of a group members' joint commitment to lie (a potential Gilbertian understanding of a group lie). She argues that neither of these accounts of a group lie are satisfactory. Instead she puts forward a view that focusses on what the group itself has said, and which therefore allows that a group can lie when no individual group member has lied, or even that a group can fail to lie when every individual group member does (Lackey, 2018). In doing so she sees her view as providing a framework for holding groups themselves responsible for acts like lying, and I regard her account as notable for this reason. Very briefly, Lackey (echoing Gilbert's criticisms of summative views of group belief in Chapter 3) thinks that a summative understanding of group lies is inadequate.[50] Firstly, this is because she suggests we can imagine cases in which all operative group members can lie to someone without the group itself having told a lie.[51] Further, Lackey also notes that there are certain situations in which we would think that a group has lied even if no operative group members have lied. She asks us to imagine a situation in which the group's spokesperson, who is not actually a member of the group,[52] offers up information on behalf of the group. Lackey suggests that the spokesperson could be subtly misled by the group into thinking that the group believed *p* without any group member actually telling the spokesperson the group believes *p*. If this were the case, the spokesperson

[48] Intending to mislead would amount to having a belief that p, but a desire to bring about the belief that ~p in someone else. We have already encountered arguments in favour of the view that groups are capable of beliefs etc. (Gilbert, List and Pettit, French).

[49] The only explicit discussion of the group lies that I have found.

[50] And here she makes it clear that both simple summativism (whereby a group can be said to X if all members of the group X) and also more sophisticated summative views (whereby a group can be said to X if all or a majority of specified operative group members X) fail (Lackey, 2018).

[51] Noting that context is important in this respect, Lackey suggests the following: "If the CEO of Philip Morris testifies that smoking does not pose health risks, whether it is the company's statement depends on where and to whom it was offered. It is one thing to state it at a board meeting to co-workers and quite another to say this to one's spouse while on vacation." (Lackey, 2018, p266). We should be disinclined to consider an institution as having told a lie where all operative group members lie to a particular person about something, but do so in their personal capacity as 'husband, best friend and son' (Lackey, 2018, p266) rather than in their professional or official capacity as a member of a particular corporation. Given the capacity for 'personal lies' (as Lackey calls them), a summative account of group lies fails.

[52] Lawyers, for example, often talk on behalf of their clients – thus it is possible for the spokesperson of a group not to be an actual member of the group.

would not be telling a lie when they said that the group believed *p*, as this would be what they believed to be the case. The group members who do not say anything to the public regarding *p* would also not have lied. Yet Lackey thinks we would nevertheless consider this a case of group lying (Lackey, 2018, p267) – and I think she is right – even though, according to a summative picture, no operative group members lied about what the group believes. It may well be the case that, if one adopts the broader definition of a lie, then group members have lied to the spokesperson – lies of omission, perhaps.[53] The group could then perhaps be regarded as having lied to the public via lying to their spokesperson. Nevertheless, even if one does favour the broader definition of a lie, the former situation in which all members of a group can lie without the group itself lying still presents a problem for a summative account of group lies.

Lackey goes on to suggest that where summative accounts of group phenomena like lying clearly fall short, one may think that accounts like Gilbert's which call for joint commitment may work. As we have seen, understanding group beliefs as involving a joint commitment separates the group belief from the beliefs of individual group members – again, so much so that a group can believe *p* by forming a joint commitment to believe *p* even when no individual group member believes *p*. As Lackey notes, this kind of view can easily account for why we think that a group hasn't lied even if all operative members lie (i.e., when they lie in their personal capacities) – namely because there is no joint commitment to tell a lie as a group. Moreover, Lackey seems to think that it can also account for why we would think that the group in question has lied if its spokesperson lies – if they jointly committed to mislead the spokesperson who would deliver the group's view to the public, we can say that there was also a deliberate attempt by the group to mislead the public (Lackey, 2018, p265).[54]

Yet Lackey argues that a joint commitment view of group lies is nevertheless still inadequate because she thinks that operative members of a group can jointly commit to lie, but still fail to lie as a group, and also that certain situations can arise where we would still think groups had lied even if there was not a joint commitment to lie (Lackey, p270). With regard to the former she asks us to imagine cases where all operative members of a group believe, for example, that there is no link between smoking and lung cancer, but keep that belief private (for whatever reason). If all the operative members of the group jointly agree to lie to the public and announce that smoking is safe – that there is no link between smoking

[53] Lackey herself does not endorse this broad definition.
[54] This also makes it easier to differentiate the group's lying to the spokesperson from certain individual group member's (who may not the authority to speak on behalf of the group) lying to the spokesperson.

and lung cancer – Lackey suggests we would struggle to make sense of this being an instance of group lying. The thought seems to be that one can form a joint commitment to do certain things, to lie, for example, without actually succeeding in lying. Indeed, she argues that "if the public learned all of the details of the case, they might regard Philip Morris as ignorant and misinformed for believing that smoking is safe, and they might even regard it as deceitful—given the members' agreement to lie—but they wouldn't say that the company lied" (Lackey, 2018, p270). What Lackey seems to be claiming here is that the public, if it learnt all the facts about the case, would not regard the group as lying in this instance because what the group 'really' believed was that smoking *is* safe, given that all the operative members of the group actually believe this.

Yet if one takes Gilbert's account of group belief seriously, one recognises that what the group 'really' believes is not necessarily whatever the members of the group believe. As Gilbert argues, it is possible for a group to have a belief that no individual group member has, and List and Pettit further detail how this could be achieved via the internal decision-making structures of an institution. Here, the belief that the company needs to have in order to lie in saying that smoking is safe is the belief that smoking is *not* safe. Yet as discussed previously as well, some beliefs that groups and institutions have may be implicit, or implicitly accepted. Yet if the group or institution explicitly agrees to tell a lie in saying for example that smoking is safe, it must clearly hold the belief that smoking is not safe – or else it would not have needed to explicitly commit to telling a lie here. That every member does actually think that smoking is safe is irrelevant. In agreeing to lie, rather than to just make an announcement, say, they must be in agreement too about what they as a group or institution take to be the actual case – that smoking is actually not safe. Again, if Lackey is correct in saying that the public would not think the company had lied in this kind of a case, it seems to me that this would only be because the public has a tendency to think about group beliefs in (what have been shown to be inadequate) summative terms.

In terms of the latter criticism Lackey says that we can tweak the example of the manipulated spokesperson in such a way that there is no joint commitment to lie as a group, but where we still think that the group has lied. If the board of the corporation holds a meeting with the spokesperson to establish what the view of the corporation is on a certain matter, and each operative member misled or lied to the spokesperson without any joint commitment to do so, Lackey argues that we would still want to see this as an instance of the corporation as having lied. I think Lackey is right that our instinctive response to the situation would be to regard the corporation is this example as having lied. Yet one suggestion for why

this is the case can be cashed out in terms of the point above that Fricker made – many kinds of groups that interact with or provide information to the public will have a kind of implicit joint commitment to trustworthiness, that 'it comes with the territory'. This situation, then, does not provide a counter-example to an account of group lying that makes use of the notion of joint commitment. Although an explicit joint commitment to lie may be missing, it would nevertheless still be the case that operative members of the group, in putting forward a dishonest representation of the view of the group were violating a joint commitment to be truthful, or trustworthy – norms we associate with the making of assertions. If we understand a lie as a violation of the norms of assertion, a group which behaved like this has lied.

Lackey however does not seem to think this is a possibility as she states that joint intentions "necessarily requires intentional activity" (2018, p270) on the part of members – i.e., are formed explicitly – and so it would be impossible for an account of group lies modelled on a joint commitment to make sense of a group telling a lie without their explicitly forming a commitment to lie. I have suggested that this need not be the case, as Gilbert herself notes that we are sometimes justified in ascribing beliefs to a group without group members even being aware of said beliefs – given that we are in certain instances justified in presuming a background joint commitment on the part of said group members to abide by what certain sectors of that group put forward as the group's views (Gilbert, 1987, p200). Indeed, since it is widely accepted that individuals have unconscious beliefs, we could also expect institutions to have beliefs they weren't aware of. Indeed, we saw in the previous chapter that institutions might not realise that they are in fact committed to a particular belief because it is rationally entailed by other beliefs they hold.

In addition, I have also argued that in institutional groups in particular it will often be practically impossible for all the members to be aware of all the things that that institution believes. As noted, it is a feature of institutions that are hierarchical (or at least have some division of labour), with different parts of the institution carrying out different functions and roles. In certain cases, some information within an institution is classified. As a result, some members of an institution are not aware of certain information (i.e., when this is not pertinent to their roles), even though this information could influence what the institution itself believes. The same would be true of institutions which have beliefs about certain subjects the content of which is itself beyond the comprehension of many members of that institution, simply because there are certain units within that institution that carry out very specialised or technical roles. Thus it is easy to imagine that a spokesperson (internal or external to the

institution) could fail to be privy to all the information the company has at its disposal, and so they could be unaware that they are telling a lie on the company's behalf.

Before moving on it is worth examining Lackey's own account of group lies. According to Lackey:

> A group, G, lies to B if and only if (1) G states that *p* to B, (2) G believes that *p* is false, and (3) G intends to be deceptive to B with respect to whether *p* in stating that *p*. (Lackey, 2018, p271)

She suggests that the above account can accommodate the suggestion that a group lie can come apart from individual group member's lies, "because what a group says can come apart from what its individual members say" (Lackey, 2018, p262). Indeed, a key part of Lackey's account of a group lie involves the group stating something, or making an assertion. Lackey suggests that a group can make an assertion either via co-ordinated effort, e.g., by co-authoring a publication, say, where they literally write it together, or via an authorised spokesperson. In the former case, a group would lie if it deliberately asserted something false with the intention to mislead when co-authoring that essay. Lackey however takes the latter to be the far more common way for groups to make assertions, and it is in these situations that the lie of the group can be clearly distinguished from the lies of individual group members. Whether or not someone has the requisite authority to assert something on behalf of the group is important for seeing whether or not what they say (or do) is representative of the group's own views. This opens up the possibility of a group failing to lie even though every individual group member does (which we saw was a potential problem for a summative account of group lies) – i.e., even if members of an institution lie about the view of their institution, if they are not in a context in which what they say is representative of the institution's views then the institution will not have lied – rather they will have (potentially)[55] lied in their personal capacities.

Lackey also points out that when a group makes assertions via a spokesperson, sometimes this spokesperson will be a member of the actual group, but sometimes they may not be a group member – as when a group gets a lawyer to speak on its behalf and represent its interests. Furthermore, Lackey notes that often spokespersons are given a certain amount of free reign with regard to what they say, as they often find themselves in situations where they have to offer up the views of the group on the fly, as is the case when being interviewed,

[55] If, of course, they also personally believe the information there are imparting is false.

for example, and so they might end up giving responses the content of which has not actually been discussed or agreed upon by the group. Nevertheless, if the spokesperson has been granted the authority to speak on the behalf of the group, what they say is what the group says. If they are good spokespersons, they will probably get these unprepared views roughly right by inferring what the group would think based on everything else they know about what the group thinks. Nevertheless, we are left with the result that, via an authorised spokesperson, perhaps who is not an actual group member, "a group can assert a proposition the content of which the individual members are wholly unaware" (Lackey, 2018, p274). As such, the group could conceivably tell a lie via their spokesperson even though no individual group member does – and even when there is no joint commitment to do so (implicit or explicit). While it seems likely that in most situations in which this happens the group will retract whatever it is that the spokesperson has said (and perhaps fire them), this is still technically a possibility – and if the spokesperson has expertise of which the group itself lacks (i.e., the spokesperson is their lawyer), the group may not know enough to retract the lie the spokesperson has told on their behalf.

With the above in mind, I think that a joint commitment account of group lying is still a very promising way to make sense of the phenomenon. As I argued above, I am not convinced that Lackey's account of the manipulated spokesperson is a problem for a joint commitment account of group lies. Moreover, it seems that an important part of deciding who has the authority to speak on behalf of the group will involve some kind of joint acceptance on behalf of the group in question on the roles and responsibilities of certain members within the institutional structure. Even in an extreme situation in which a group is unaware that a lie has been told by an external spokesperson on their behalf, I suspect that the public's response would be to regard this lie as endorsed by the group insofar as they endorsed the spokesperson. On the other hand, if the public were to find out that the group, lacking the requisite expertise to comprehend what was said on their behalf, had let the lie of the spokesperson (with the spokesperson maintaining their position as spokesperson) stand unchallenged out of ignorance, I am not convinced they would regard this as the lie of the group. Indeed, they may well see it as the lie of a cunning and manipulative individual who held the title of spokesperson. Nevertheless, Lackey's central claim that one is able to separate what a group says from what its members say, and the importance of being able to do so, is extremely useful in terms of being able to distinguish a group lie from individual lies.

Self-deception is frequently explained in terms of ‘lying to oneself’. In much the same way that one can separate the lies of institutions from the lies of individuals, I now wish to argue that one can similarly separate the self-deception of an institution from the self-deception of individuals. This matter will be slightly complicated by the possibility of separating the notions of lying to oneself and self-deception. In the forthcoming chapter I will argue that institutions themselves are capable of both lying to themselves and of being self-deceived.

Chapter 5: Institutions can be Self-deceived and Lie to Themselves

> *"If you want to assert a truth, first make sure it's not just an opinion that you desperately want to be true." - Neil deGrasse Tyson*

We have seen that there are many examples of institutions that espouse certain views, or perform certain acts, that have, to varying degrees, a negative impact on their societies. We have also seen that in some situations, the institutions responsible mislead the public regarding the nature or extent of the wrongdoing (e.g., when BP failed inform us how much oil was really spilled in the Gulf of Mexico), or deny that they had intentionally done anything wrong at all (e.g., when Spanghero denied knowingly selling horse meat as beef). One possible explanation is that in such situations these institutions are simply lying. However, another possible explanation is that these institutions are self-deceived, or lying to themselves. Recently it has been suggested that self-deception is something that can affect institutions as well as individuals (Trivers, 2000; Deweese-Boyd *forthcoming*). In this chapter I will argue that this is indeed possible.

However, almost every aspect of self-deception is disputed in the literature – from its very definition (Mele 2001), to whether or not one is responsible for one's self-deceptive beliefs (Mele 2001, Levy 2004, Nelkin 2012, Galeotti 2016), and even whether or not being self-deceived is a bad thing (Taylor 1989, Kirsch 2005, van Leeuwen 2009). Thus I will begin this chapter by examining some of the problems that surround offering an account of self-deception in Section 1, and in doing so I draw a distinction between self-deception and lying to oneself – terms which are sometimes used interchangeably. In Sections 2 and 3 I will then argue that institutions are capable of both self-deception and of lying to themselves.

At the beginning of Chapter 4 we saw that lying has negative consequences not only for those lied to, but also for the liars themselves. The price for self-deception or lying to oneself could thus be two-fold. In Chapter 6 I will discuss the moral implications of institutions being self-deceived and lying to themselves. I will also look more specifically at what institutions can do to avoid these two states.

Section 1: Self-deception and Lying to Oneself – A Distinction

There are many instances in the literature on self-deception in which the terms 'self-deception' and 'lying to oneself' are used interchangeably, or understood in terms of one

another – for example, according to Raphael Demos, "Self-deception exists […] when a person lies to himself" (Demos, 1960, p588). As indicated above, formulating a comprehensive philosophical account of self-deception has proven extremely challenging, in part because the notion of self-deception has been historically understood in a very literal way. If one models self-deception on ordinary cases of interpersonal deception it would appear that in order for someone to be *self*-deceived they must take on the role of both the deceiver and the deceived, which seems to be paradoxical. Thus, someone ought to be aware of the truth about something (i.e., that *p* is true) in order to be able to lie about it. Yet, they must also fail to know the truth in order to be taken in by the lie (i.e., come to believe that *p* is false). The self is both the liar and the one lied to – a picture which would seem to require a person to simultaneously hold two contradictory beliefs at once. Such a state is referred to in the literature as the 'static paradox'. Yet the process of deceiving oneself is also thought to be highly paradoxical – it just doesn't seem as if one's deliberate efforts to deceive oneself could ever be successful (Galeotti, 2012, p42-43). This problem is referred to as the 'dynamic paradox', and has to do with the seeming impossibility of ever successfully deceiving oneself into believing ~*p* while believing *p*.

There have been attempts to resolve these paradoxes by introducing notions of temporal and psychological partitioning. Temporal partitioning could involve, for example, a sufficient amount of time passing between someone intentionally keeping a doctored diary, and then coming to believe the falsehoods written there once they have forgotten that they doctored the diary (Mele, 2001, p16). While this is clearly an instance of intentionally deceiving oneself, these sorts of cases are fairly unusual, and not the kind of case people usually have in mind when talking about self-deception.

Psychological partitioning on the other hand often involves positing some kind of division of the self, where one part of the self is seen as the deceiver and another part of the self the deceived. For example, Roy Sorensen notes, "one popular way to disarm the contradiction [of self-deception] is to divide the self into parts, homunculi, and say one homunculus is deceiving another homunculus" (Sorensen, 2005, p343). In less extreme versions, the psychological partitions involved may only be between the contrary beliefs *p* and ~*p*, with the contrary beliefs occupying "semi-independent departments within the same mind" (Davidson, 1982, p149). Yet these attempts to avoid the paradoxes cited above by dividing the self or the mind have in turn been criticised for either failing to cover a wide

enough range of cases we would ordinarily see as instances of self-deception, or for introducing further puzzles of their own (Davidson, 1982, p140; Mele, 2001).[56]

Some recent philosophical work on self-deception has tended to adopt a 'deflationary' approach to the phenomenon.[57] These deflationary accounts side-step the apparent paradoxes mentioned above by suggesting that self-deception is generated through an *un*intentional process (that lies beyond conscious awareness). Given the intentional nature of lying, it is possible to draw a distinction between self-deception so understood, and 'lying to oneself' – a description which would only fit the explanations of self-deception that are modelled on interpersonal deception and appeal to temporal or psychological partitioning. If, however, one is sympathetic to the criticisms directed at these accounts of self-deception which describe it as involving intentionally lying to oneself, one may well think that individuals are (virtually)[58] incapable of lying to themselves, but could still be susceptible to self-deception – if one adopted a deflationary account (institutions on the other hand would appear to be perfect candidates for lying to oneself, as they are made up of a number of individual members, some of whom could lie to one another – see Section 3 below).

On deflationary accounts of self-deception, the self-deceived subject will not be *actively* trying to bring about a false belief, and consequently the subject may not be aware of the process whereby they procure their self-deceptive belief. The self-deceived subject will also only have one, false belief, despite technically having access to evidence that strongly suggests the opposite is true. This gives rise to a situation in which the subject will not literally have to juggle two incompatible beliefs, p and $\sim p$ – they just have access to evidence which ought to lead them to believe $\sim p$ instead of p. An account along these lines has been

[56] While Davidson is an advocate of understanding self-deception in terms of mental partitioning, he does note some concerns that have been raised against this view, including worries that the view seems to require us to make sense of attributing things like beliefs and desires and even actions to something less than a whole person, as well as concerns along the lines suggested by Sartre that "the notion of responsibility would lose its essential point if acts and intentions were pried loose from people and attached instead to semi-autonomous parts of the mind" (Davidson, 1982, p140).

[57] One criticism that has been launched against deflationary accounts is that they cannot adequately distinguish between cases of self-deception and cases of wishful thinking. For those who subscribe to an intentional model of self-deception, the intention is what differentiates self-deception from wishful thinking (Mele, 2001, p73; Szabados, 1973, p203). Yet those who put forward deflationary accounts suggest that the difference between self-deception and wishful thinking may lie in the strength of the evidence against the belief – wishful thinkers do not encounter strong evidence against what they wish to be true, while self-deceivers persist in believing what they wish to be true in the face of strong evidence to the contrary (Szabados, 1973, p204-5). Alfred Mele has also suggested that if we are talking about wishful *thinking* rather than wishful *believing*, this could be another distinguishing feature of self-deception which involves beliefs (Mele, 2001, p73). Exactly how we are to distinguish wishful thinking from self-deception, given the contested understanding of self-deception, is an issue which is up for debate.

[58] I am sympathetic to the temporal partitioning view of lying to oneself – I just don't think such cases are at all frequent.

put forward by Alfred Mele. One of the strengths of his account is that it can successfully side-step the paradoxes previously mentioned without introducing any complex "mental exotica" (Mele, 2001, p4) to explain them away. Instead Mele references a number of well-documented psychological phenomena (e.g., negative and positive misinterpretation, selective focusing, confirmation bias etc.),[59] thus rendering self-deception non-mysterious as well as non-paradoxical. On Mele's view, a subject S is self-deceived if:

> 1. The belief that *p* which *S* acquires is false.
> 2. *S* treats data relevant, or at least seemingly relevant, to the truth value of *p* in a motivationally biased way.
> 3. This biased treatment is a nondeviant cause of *S*'s acquiring the belief that *p*.
> 4. The body of data possessed by *S* at the time provides greater warrant for ~p than for *p*.
> (Mele, 2001, p50-51)

Thus, according to Mele, self-deception involves having only one false belief. The response of *S* to what is claimed to be 'data' or evidence for (or against) *p* will be biased as the false belief is supported by unconscious desires or emotions, and motivationally biased evidence gathering (Mele, 2012, p2-3). The biased treatment of evidence for (or against) *p* will be what leads to the belief that *p* (i.e., it will not be an accident that, given their handling and interpretation of the evidence available, they come to believe that *p*). As a consequence, even in the face of evidence to the contrary, by selectively attending to their surroundings a person may come to believe *p*, given their own desire for *p* to be true (Mele, 2012, p2-3).

Thus, for example, we can imagine a situation involving a young man, Don, who is a very heavy drinker. When his friends attempt to bring up the possibility that he might have a drinking problem, Don gets very angry and defensive, stating he does not think he has a drinking problem. He thinks this even though there's plenty of evidence to suggest that he is, in fact, an alcoholic – e.g., his home is full of empty beer bottles, and he has repeatedly been sent home from work for drinking on the job, or arriving at work drunk. Don can meet Mele's four conditions, so there is reason to think that Don's assertions that he is not an alcoholic are the result of self-deception. In other words he 1) falsely believes that he is not an alcoholic, when he in fact is, and 2) he treats data relevant, or at least seemingly relevant, to the truth

[59] Mele cites negative misinterpretation (the desire that p leads us to discount evidence against p, something we wouldn't do in the desires absence); positive misinterpretation (the desire that p leads us to count evidence for ~p as evidence for p, something we wouldn't do in the desires absence); selective focussing/attending (where the desire that p focuses our attention on evidence for p, rather than on evidence that goes against p); and selective evidence gathering (where our desire that p enables us to find difficult to come by evidence for p, while overlooking easily accessible for ~p). Mele goes on to talk about how these psychological tendencies can combine with others, like confirmation bias, or how vivid information appears to us (Mele, 2012).

value of this belief in a motivationally biased way – for example by comparing the amount he drinks to that of his alcoholic uncle rather than to a more objective standard, or perhaps he thinks that his boss has some personal vendetta against him, and that the accusations of being drunk, or drinking, at work are thus unfair and untrue. Moreover, 3) this biased treatment of evidence is what leads to his acquiring the belief that he is not an alcoholic, and 4) Don is aware of amount he drinks, he is aware of the excessive amount of empty bottles in his house, and aware of the reasons his boss cites when he sends him home, and thus the body of data possessed by Don at the time provides greater warrant for believing that he is an alcoholic than that he is not. In this respect, even though Don denies being an alcoholic, his friends and family who are also aware of all this evidence think he has a drinking problem, and any third party who became familiar with all this information would conclude that his denials must be the result of self-deception.

If his family and friends did not attribute self-deception to Don, perhaps they would think Don was just lying when he said he did not have a drinking problem. Here it should be noted that there can some difficulty in accurately identifying cases of self-deception. Indeed, in the above example it is possible that Don is well aware that he has a drinking problem, and is simply lying to friends and family when he says that he does not. He may do this because he doesn't want to have uncomfortable conversations with those around him and be obliged to make any changes to his lifestyle, or he may lie because he would be ashamed to admit that he's an alcoholic. Yet if Don is in fact genuinely self-deceived about his drinking problem, he may behave in the exact same way – for example, if he is lying he will deny that he has a drinking problem when confronted because he wishes to mislead those around him regarding his alcoholism, but if he is self-deceived he will also deny that he has a drinking problem when confronted because he does not believe that he does. When confronted with someone who avows some or other thing, while having access to evidence to the contrary, we may not be able to establish without doubt whether or not they are lying to those around them, or are self-deceived.[60] Nevertheless, if we know enough about Don and his situation (that he should be classified as an alcoholic, what desires related to the matter he may have, the evidence available to him, perhaps whether or not he is typically an honest person), we can form a view regarding whether or not he can meet Mele's four criteria. If we determine

[60] The use of the phrase 'lying to themselves' would have more of a stylistic ring to it, but given the distinction made above between 'self-deception' and 'lying to oneself', if Don meets Mele's criteria, he is self-deceived.

he can, then Don is a candidate for self-deception.[61] Moreover, in spite of similar challenges, I will argue that the same is true for institutions.

Section 2: Institutions can be self-deceived

I will return now to the case of the Ford Pinto mentioned at the beginning of Chapter 4. Three women died in a tragic motor accident in 1978 in which a van crashed into the rear end of their car, a Ford Pinto, causing the gas tank to rupture, and the car to burst into flames. Following this incident Ford was charged with three counts of reckless homicide, the first ever such charge to be brought against an American corporation (Hoffman et al., 2014, p139). This was not the first fatal accident involving a Ford Pinto. Just prior to the 1978 accident, Mark Dowie published an article entitled 'Pinto Madness,' which appeared in the popular investigative magazine *Mother Jones*, and directly accused Ford of the following: a) rushing production, b) becoming aware during tests that rear-end collisions caused the fuel tank to rupture extremely easily, c) continuing with production despite this discovery, because the machinery on the assembly line was already set up to implement that design (and it would cost a lot of money to reconfigure the machinery on the assembly line), and d) lobbying for more than eight years against government safety regulations that would have forced them to recall the Pinto as it would not have met the proposed safety standards (Dowie, 1977, p18).

Ford disputed the above, saying that the Pinto was not unsafe, that the number of deaths linked to the exploding fuel tank was statistically irrelevant, and that the Pinto had in fact met or surpassed the safety standards set by the government, and had recalled vehicles as soon as ordered to do so (Hoffman et al., 2014, p140). In his closing statement at the trial, Ford's attorney further noted that Ford employees had bought the Pinto for themselves and their families, so they clearly must have thought the cars were sufficiently safe (Hoffman et al., 2014, p142). It was also pointed out that the accident in question occurred because the three women stopped their Ford Pinto on a narrow highway – that the car was not moving at the time of impact meant that, given the speed at which the truck collided with it was going, any other similar sized car would have suffered the same kind of damage (Hoffman et al., 2014, p143; Fisse and Braithwaite, 1987, p246-7). Ford was eventually found not guilty of criminal homicide.

[61] And we do in fact frequently diagnose others with self-deception. That we are rarely in a position to be absolutely confident in our diagnoses, citing self-deception is often the best explanation we have for the behaviour of others.

However, it also turned out that the court had ruled that Ford's test results pre-1973 were not admissible in court, so although certain test results indicated that Ford knew about the exploding gas tanks in their Pintos "as early as 1971" (Hoffman et al., 2014, p143), this was not taken into consideration. It is also the case that "the company was aware, because of tests with the Capri, that the over-the-axle position of the gas tank was much safer than mounting it behind the axle. [Yet] Ford decided to mount it behind the axle in the Pinto to provide more trunk space and to save money" (Hoffman et al., 2014, p143).

I want to make use of the above case to determine whether or not an institution is capable of meeting Mele's four minimal conditions for self-deception. If Ford can meet Mele's minimal conditions for self-deception, there is reason to think that Ford was self-deceived about the Pinto. The details of the Ford Pinto case I outline here are significantly simplified – more details can be found in the original trial transcript "State of Indiana v. Ford Motor Company" and in the article "State of Indiana V. Ford Motor Company revisited" by Becker et al. (2002). Ford Motor Company has a belief that *p* ('the Pinto is a safe car')[62] and a desire (to break into the small car market and make a profit from selling their smaller cars as soon as possible). Can we now say that Ford was self-deceived according to Mele? I will assume that the belief that *p* (that 'the Pinto is a safe car') espoused by the company is in fact false, and therefore that Ford can meet Mele's first condition ('the belief that *p* which S acquires is false').

Mele's second condition is 'S treats data relevant, or at least seemingly relevant, to the truth value of *p* in a motivationally biased way'. There is plenty of evidence to suggest that the Pinto was not in fact a safe car – as mentioned above, performance tests in 1971 indicated that the fuel tank would rupture and the car would burst into flames if the vehicle suffered a rear-end impact. Ford was also aware that placing the fuel tank in different positions didn't yield this result. Yet Ford went ahead with manufacturing and selling the Pinto, with its dangerous fuel tank placement, which suggests that Ford held the belief that the car was sufficiently safe. Indeed, as noted, one of the factors which appeared to count in Ford's favour in the court case was the fact that Ford employees, including engineers, even

[62] As a reminder, Gilbert (2014) suggests that for a group to have a belief, members of that group need to jointly commit to believe *p* as a body (which involves emulating as far as possible a single body that believes *p*). Note that she insists that this group belief is not reducible to individual beliefs, because the group's overall belief can be at variance with what its individual members believe – and moreover, she does not think that this alters the 'truth-tracking' and non-volitional aspect of such beliefs (in part at least because she views them as constrained in much the same way that Peter French does when he cites Peter Drucker as saying "the corporation […] must have a basic policy. For it must subordinate individual ambitions and decisions to the needs of the corporation's welfare and survival. That means that it must have a set of principles and a rule of conduct which limit and direct individual actions and behavior" [French, 1979, p213]).

bought the vehicle themselves. A Ford engineer (who wished to remain nameless) is quoted in the *Mother Jones* article as saying, "this company is run by salesmen, not engineers; so the priority is styling, not safety" (Dowie, 1977, p23). It is plausible then that perhaps what was going on was that Ford's desire to release the car and start making money as soon as possible made Ford give more weight to test results which supported this desire, while test results which inhibited this desire were less salient to Ford. As Mele notes:

> [M]otivation can increase the vividness or salience of certain data. Data that count in favor of the truth of a hypothesis that one would like to be true might be rendered more vivid or salient given one's recognition that they so count; and vivid or salient data, given that they are more likely to be recognized and recalled, tend to be more "available" than pallid counterparts. (Mele, 2001, p29)

Thus it seems that 1970s Ford may meet Mele's second condition. Although Ford technically had access to evidence which should have led them to believe that the Pinto was not sufficiently safe, due to motivationally biased evidence gathering this evidence was not as salient as the more minimal evidence available that lent support to the preferred view that the Pinto was safe. Mele's third condition is that 'this biased treatment is a nondeviant cause of S's acquiring the belief that p'. Ford's disregard of the test results which indicated that the Pinto was unsafe, something which may have been influenced by their desire to release the Pinto as soon as possible (this was a company 'run by salesmen' after all), led to their belief that the Pinto was sufficiently safe.

Finally, Mele's fourth condition is that 'the body of data possessed by S at the time provides greater warrant for ~*p* than for *p*'. This can be quite complicated to ascertain, especially as institutions like corporations are hierarchical in nature. Thus, while certain members or departments of the organisation may have had access to information or evidence which provided greater warrant for the belief that the Pinto was not safe (i.e., the team that performed the safety tests in 1971), this evidence may not have been made available to the rest of the company, or to the relevant members of the company who set Ford's agenda (the board, or the CEO). There are two possible scenarios here – the first being that relevant decisioning-making members of Ford were not aware of the test results, and a second scenario in which they were – and I suggest that in both cases Ford would count as self-deceived.

In the first scenario we can imagine that a number of safety tests were performed on the Pinto – and while many of these tests yielded worrying results regarding the Pinto's

propensity to burst into flames upon sustaining a rear-end impact, some tests where certain conditions were tweaked slightly in some or other way yielded more favourable results. It could easily be the case that the team running the safety tests, alert to Ford's agenda of releasing the Pinto as soon as possible, gave more credence themselves to the test results that supported the view that the Pinto was safe, and dismissed for a number of reasons test results that indicated the opposite. Given this, we can imagine that the team involved in running the safety tests would have passed along a positive report on the Ford Pinto's performance (keeping the unfavourable test results filed away in a cabinet somewhere, perhaps, regarding them as details that were not pertinent to the 'bigger picture'), which would have influenced (in a nondeviant way) Ford's belief and subsequent decisions. Nevertheless, Ford possesses information that provides greater warrant for the belief that the Pinto was *not* in fact safe, rather than that it was. That this data was not taken to be pertinent by Ford is another matter (and precisely the issue).

Mele himself notes that he does not see his fourth condition for self-deception as strictly *necessary* – because he thinks that it can be the case that one can be self-deceived in believing *p*, where the evidence available favours ~*p* rather than *p*, but for it to be the case that the evidence one actually possesses does in fact favour *p*, due to selective evidence gathering (Mele, 2012, p3). As I have described scenario 1, Mele's fourth condition for self-deception is met only by the team performing the safety tests. From the viewpoint of the CEO and other decision-makers, condition four is not met – the evidence before them justifies Ford's belief that the Pinto was sufficiently safe. Yet Ford will, in scenario 1, still count as self-deceived even if one wanted to limit the view of Ford to just those members who are responsible for espousing the view of the company (the board of directors, say) who we have imagined do not know about all the available information regarding the safety results. That they do not possess these results themselves would be, much as Mele suggests is possible, due to the selective evidence gathering that occurred in the testing department due to Ford's desire to release the Pinto as soon as possible. The desire of Ford to release the Pinto as soon as possible, a desire of which the testing department was aware, influenced the selection of certain evidence, and caused other evidence to be overlooked.

Yet even if, in a second scenario, every relevant member of the company *were* aware of the existence of these test results – the entire set of test results (both the few positive, and the negative ones) had been handed to the board of directors, the desire of the company to release the Pinto as soon as possible may have made the decision-makers of Ford give less weight to this information. As a result, those directly responsible for Ford's beliefs and

actions may simply have failed to recognise that the Pinto was in fact unsafe in spite of having access to a body of data which should have supported the view that the Pinto was in fact unsafe. As such, it seems that Ford is also able to meet Mele's fourth criterion in scenario 2, that 'the body of data possessed by S at the time provides greater warrant for ~*p* than for *p*'. That Ford can plausibly meet all four of Mele's conditions outlined above gives us reason to conclude that Ford – and any number of institutions – is a strong candidate for self-deception.

Robert Trivers offers a reason for thinking that self-deception would be common in an institutional setting. In his paper, "The Elements of a Scientific Theory of Self-Deception", Trivers puts forward an evolutionary explanation for self-deception, suggesting that self-deception may have come about so as to better deceive others (2000, p115). Trivers goes on to suggest that something similar can happen in organisations, saying that "when an organization practices deception toward the larger society, this may induce organizational self-deception" (Trivers, 2000, p126). He argues that organisational self-deception within NASA led to the Challenger disaster, at least in part because NASA was under pressure to secure support for further missions into space. After the first successful moon landing, there was less urgency and less interest from the public for such projects (2000, p127). As a result, NASA had to make their proposed projects look as uncomplicated and financially viable as possible in order to continue to exist as an organisation (2000, p128). Thus, when the Challenger mission was put forward, the sector in charge of safety was under enormous pressure to downplay the risks involved, and any problems with the design of the spacecraft, as the organisation desperately needed to secure funding from the government and public support for the project (2000, p126). Consequently, the safety unit in NASA failed to accurately convey the dangers associated with the O-ring design (the fact that the O-rings sustained alarming degrees of damage in low temperatures) – even reinterpreting this design weaknesses as a strength (2000, p128). As a result, Trivers suggests that NASA suffered from self-deception:

> In short, in service of the larger institutional deceit and self-deception, the Safety Unit was thoroughly corrupted to serve propaganda ends, that is, to create the appearance of safety where there was none.
> There is thus a close analogy between self-deception within an individual and self-deception within an organization, both serving to deceive others. (Trivers, 2000, p128).

Parallels can be seen between this case and that of the Ford Pinto.[63] That Ford was under pressure to sell the Pinto – and perhaps under more pressure than usual as they were branching out into the small car market for the first time with the advent of the Pinto – would have had an effect on how the company wanted to portray the car to the public, and perhaps what it actually believed about the Pinto.

Moreover, institutions may well be particularly prone to self-deception. As they are often made up of hierarchical, specialised subgroups situations may frequently arise when certain subgroups are just not privy to information. This could easily lead to situations like scenario 1 described above. Even in situations in which information is freely disseminated between all sectors of an institution it may be the case that different departments, or those who have to make the final decision on some or other matter, lack the specialised knowledge to grasp certain important nuances of the information being shared, and thus they may often have to rely on the opinions of others, which could be unreliable or skewered because of the institution's desires and goals. Additionally, as institutions are goal-orientated – they are established for a reason – they are often answerable to shareholders or the public. As a result there can often be a lot of pressure in institutional settings to produce the 'right' results.[64] Trivers also points out that, within a group or institutional setting, self-deception may be especially difficult to detect and root out, because "within each group individuals are misoriented in the same direction, easily reinforcing each other and absence of contrary views is taken as confirming evidence (even silence is misinterpreted as support)" (Trivers, 2000, p129).

Ian Deweese-Boyd has also discussed the possibility of groups being self-deceived. For Deweese-Boyd, what distinguishes what he calls collective self-deception from individual self-deception is simply its external, social setting (*forthcoming*, p3-4). He

[63] While any deception of the public in the case of the Ford Pinto may not have explicitly been part of the institutional agenda (although we saw in the previous chapter that institutions can lie, and seem to do so fairly regularly), as noted above, Ford was a company 'run by salesman', and one of the chief tools of salesmen – advertising – has long been criticised as manipulative and (at least sometimes, and to some degree) deceptive (Sharma and Chander, 2011, p191). Indeed, Gary Armstrong, et al. open their 1979 article entitled "Detecting and Correcting Deceptive Advertising" by pointing out that "the subject of deceptive advertising has been discussed in law journals for over half a century" (1979, p237). Clearly, then, this is a long-acknowledged phenomenon.

[64] Dan Ariely cites the example of someone who works for an economic consulting firm who noticed that a colleague of theirs who refused to overbill clients "consequently had an overall billing rate that was about 20 percent lower than the average […] when it was time to lay people off, he was the first to go" (Ariely, 2012, p36). This is illustrative of the kind of pressure that members of a company may be under to produce certain results. The company itself may feel pressured to produce a certain profitability in order to please shareholders, and this pressure could in turn affect the behaviour of certain departments, or individual employees, for example.

discusses the possibility of collective self-deception only in what could be described as a summative manner. When Deweese-Boyd talks about the collective self-deception of unreflective consumers, what he seems to be suggesting is that many people share similar biases and, as a consequence, form and maintain the same false belief in the face of strong evidence to the contrary (*forthcoming*, p3). Yet he makes a similar point to Trivers above when he discusses what he regards as the self-deception of consumers regarding their purchasing choices. For Deweese-Boyd, self-deception in a social context is often easier to develop, and more difficult to extricate oneself from (*forthcoming*, p3). He argues that this is so because a person often does not meet any opposition to their self-deceptive beliefs – everyone within that group context is similarly self-deceived. Consequently, one's self-deceptive belief is never challenged. In fact, if one does waver in one's self-deceptive belief this may bring about censure. Although he does not talk of entities like institutions being self-deceived, what he says could still be relevant to the more robust institutional self-deception I am talking about. Gilbert makes a similar observation, suggesting that "it is important to see how participation in a collective belief can have consequences for one's private thoughts, inhibiting one from pursuing spontaneous doubts about the group view, inclining one to ignore evidence that suggests the falsity of that view, and so on" (Gilbert, 2000, p44-45). In an institutional setting where there is a particular agenda or institutional culture, the same difficulty regarding recognising and escaping self-deception would likely occur when an institution itself is self-deceived. In the case of the Ford Pinto, any nagging doubts about the safety of the vehicle by any employees may well have been put aside given that the company had put forward a certain view which everyone who worked there would have been expected (perhaps after some discussion) to accept.

Section 3: Institutions also Lie to Themselves

In addition to the above whereby the institutional self-deception that occurs is unintentionally brought about, we can also imagine scenarios in which an institution intentionally lies to itself. For example, the test results which indicated that the Pinto was not safe when suffering a rear impact may have been deliberately withheld by team carrying out the safety tests due to pressure from above to release the car as soon as possible. This would mean that they were not simply treating the evidence available in a biased way and reaching the wrong conclusion – rather, they were well aware of the fact that the Pinto was unsafe and deliberately concealing this fact. Indeed, while it was noted that some engineers at Ford bought the Pinto

for themselves, indicating that they thought the car was perfectly safe, we do not know exactly which engineers these were, and whether or not *they* had access to the evidence which clearly indicated that the cars were in fact unsafe. Alternatively, it may also have been the case that the team conducting the safety tests did pass along the worrying results, and that these were suppressed by those higher up the hierarchy who wanted to go ahead with the release of the Pinto as soon as possible anyway.

In a recent paper, Holly Lawford-Smith notes that groups which are structured in a hierarchical fashion can often be complicated by what she calls 'nested agency', where "subgroups within the collective are themselves collective agents" (Lawford-Smith, 2018, p334). We saw in Section 2 above that the team conducting the safety tests in scenario 1 formed a sub-agent within Ford that was self-deceived. The biased report, or the self-deceptive beliefs, of this particular sub-agent went on to inform the beliefs of Ford as a whole. This notion of nested agency can also explain how Ford could lie to itself in our present example. If there were some employees or departments at Ford who knew that the Pinto was unsafe and deliberately misled the rest of the company about this, we could have independent sub-agents within an institution deceiving one another.

We saw above that individual self-deception modelled on interpersonal deception is paradoxical, and accounting for self-deception pushed some theorists to put forward deflationary accounts of the phenomenon. However, because institutions can literally contain different minds in virtue of the institutions' many members and specialised departments, it is conceivable that institutions could believe two contradictory propositions at once, and that an institution could quite literally lie to itself. When sub-groups within an institution lie to each other, this is in effect an example of an institution lying to itself.

Again, I think that a closer parallel with Mele's account of self-deception is also possible – for example that the agenda set by the company's desire to break into the small car market and make a profit as soon as possible could effectively blind different sectors in the company to the importance of certain test results (as outlined in scenario 1 above). However, if one begins to view what appears to be self-deception in institutions in line with theories of self-deception that were modelled on interpersonal deception and the notion of a mind containing a number of homunculi, this would mean that institutional entities are particularly susceptible to what is sometimes thought of self-deception, but which I have labelled lying to oneself. As has been noted, institutions are often highly hierarchical, containing numerous different and specialised subgroups which are responsible for different duties within the

institution and which, as a result, have access to information to which other parts of the institution may not have access.

Here one may question why we ought to be concerned with institutions at all if we begin to break institutions up into various subagents (which, conceivably, might even be made up of just one person if say, just one person within a company is responsible for some or other duty), as it begins to look as if we are not talking about institutional self-deception at all anymore – and arguably not even self-deception modelled on interpersonal deception, but rather just straightforward interpersonal deception. The response here would be that we are concerned with the actions of the institution itself. Ford was selling unsafe cars, not Bob from the safety department, or even Lee Iacocca (the president of Ford Motor Company at the time), or Henry Ford II (the then CEO), and in order for something to be seen to act it needs to have beliefs and desires. While we often do conflate those in charge of a company with the company itself, one of the key features of an institution is that its members are fungible – an institution can remain the same institution despite a turnover of its members, and thus it is a mistake to regard the actions of an institution, like Ford, as those of particular members, like Lee Iacocca.

Although one does hear of rogue members acting on their own within institutions, there is no reason to think that there was any single individual at mid-70s Ford had gone rogue. It is also perhaps helpful to consider whether or not, had Iacocca been replaced with someone else, Ford would still have gone on to manufacture and sell the Ford Pinto. Peter French's suggestion that an action can be properly attributed to a corporation if it has been "performed for corporate reasons" (French, 1979, p213) may further help to make sense of the distinction between interpersonal lying and an institution's lying to itself. Given the overarching aims of the corporation, and the mechanisms that were in place, we should be able to conclude that it was Ford selling the Ford Pinto, not any individual employee. Moreover, if the reason that the results of the safety tests on the Ford Pinto were either deliberately withheld or supressed were *corporate* reasons – i.e., in line with the company's beliefs and desires – then the company would seem to be lying to itself.

A further reason to see an institution as lying to itself is that the structure of the institution provides the conditions for what occurs. One cannot understand the phenomenon without understanding the institution, and recognising the fact that there must be something about the institution itself that provided the right conditions for what went wrong. One possibility here could be the company ethos, perhaps a view that making a profit at any cost is paramount. Such could arguably be said to be the case with Ford and its rushing the

production of the Pinto, and going ahead with producing cars that (some part of) it knew were unsafe because Ford needed to be competitive as soon as possible, and furthermore making the kinds of design decisions that it did, instead of safer ones, because these were more cost-effective.

Institutions may be structured in ways which problematically enable institutions to lie to themselves. This could be the case, for example, if there are a lack of protocols and regulations within an institution such that one section of a company, say, could go ahead with some or other action without any checks, or could pass on important information regarding some project without having to provide any supporting documents or evidence for that information. In such situations there is something wrong at the institutional level in terms of how the institution is structured, or institutional policy, such that the institution ends up with false or problematic views and acts on them. As a result, the institution and its functioning should come into focus when it comes to identifying and addressing certain problems. What may appear to be a case of an institution lying could turn out, upon closer inspection, to be something better described as self-deception, or lying to oneself. Indeed, whether an institution lies, is self-deceived, or lies to itself, it may lead to the same consequences. For example, if Ford lied to the public about the safety of the Pinto in order to break into the small car market and sell more cars, members of the public would have bought the Ford Pinto with the same disastrous outcome as in a situation where the Pinto was released because Ford was self-deceived through the safety department or the relevant decision-makers, or a situation in which some or other sub-agent within Ford lied to other parts of the company, and Ford came to falsely believe that the Pinto was in fact safe.

While it is possible that in this particular example Ford was simply lying when it insisted that the Pinto was sufficiently safe, again, it must be noted that it can be difficult to tell whether or not an avowal by an institution is an instance of straightforward institutional lying, or whether the institution is self-deceived, or lying to itself. Given that the difficulty of distinguishing lies from self-deception is also problematic in cases involving individuals, this need not be an insurmountable problem for the possibility of attributing institutional self-deception, particularly if institutions are capable of, say, meeting Mele's minimal conditions for self-deception – at least, no more problematic than it would be for attributing self-deception to any individual. An institution lying to itself may also be difficult to distinguish from a straightforward institutional lie since when an institution lies it is likely that a number of its members would also be ignorant of the truth. Nevertheless, careful attendance as to who

was meant to be deceived – the public, or another sector of the institution itself – should provide clarity in such cases.

I have argued that institutions are capable of self-deception – although in practice it may be difficult to determine when institutions are genuinely self-deceived, as opposed to lying, for example. I have argued in addition that institutions also offer a fertile setting for a very literal take on self-deception (something that, as noted above, causes trouble for an account of individual self-deception), and so it is therefore be a mistake to see all instances of what appears to be institutional self-deception as operating identically to a deflationary accounts of individual self-deception – institutions can also lie to themselves. Acknowledging that institutions are capable of self-deception and lying to themselves will have implications for how we think about the responsibility of institutions, and the ways in which responsibility ought to be distributed within such groups. What we can do to prevent these situations from occurring is also of importance. It is these topics to which I now turn in Chapter 6.

Chapter 6: Moral Implications and Possible Solutions

> "Corporations are like protean bacteria; you hit them with accountability and they mutate and change their names." - Doug Anderson

We saw in Chapter 5 that institutions are capable not only of self-deception (understood in deflationary terms), but also that institutions are able to lie to themselves without giving rise to the number of puzzles which this more literal take on self-deception raises when applied to individuals. Self-deception is generally regarded as something undesirable and problematic – it can blind one to the truth about some or other matter, make one oblivious to one's own moral shortcomings, and so lead to harmful actions and prevent one from addressing certain problems or character defects. On an institutional level, the potentially harmful consequences of self-deception are magnified, making instances of institutional self-deception cause for some alarm. Given the harms attendant on both lying and being lied to outlined in Chapter 4, it is not difficult to see that an institution's lying to itself can be also be regarded as a highly undesirable phenomenon. Indeed, I suggested in the previous chapter that the consequences of institutional lying, institutional self-deception, and an institution lying to itself can be the same. Yet when an institution causes harm, determining whether this harm was the result of an institution being self-deceived, or lying to itself, as opposed to outright lying will affect how we view that institution. In this chapter I will look at the moral implications of institutional self-deception and an institution lying to itself.

In Section 1 I begin by briefly recapping what I have argued for in Chapters 3 and 4, namely that institutions can rightly be regarded as members of a moral community, capable of praiseworthy and blameworthy actions. I put forward the claim that if we do not hold institutions themselves responsible, then we are in danger encountering what List and Pettit call a "deficit of responsibility" (List & Pettit, 2011, p165). It is also the case that, if we could hold institutions themselves responsible in a substantial way then we can minimise the risk of what P. Eddy Wilson calls "corporate scapegoating" (1993). I go on to address the concern that, in holding institutions responsible for certain wrongdoings, individuals involved in these institutional acts may escape responsibility. I argue that this need not be the case, that one can hold individuals as well as institutions responsible for institutional wrongdoings.

I will then discuss the problems surrounding self-deception and attributions of moral responsibility in Section 2. Indeed, one may worry that the possibility of institutional self-deception could absolve institutions of their unethical behaviour – and perhaps even provide

a convenient excuse that institutions could use for their unethical behaviour. I am sympathetic to views that retain some degree of responsibility for self-deception in the individual case, and so will argue that institutions need not be excused for behaving unethically when self-deceived. However, I will also claim that what may normally be seen as exculpatory in cases of individual self-deception does not necessarily apply in cases of institutional self-deception. Consequently, the success of attributions of responsibility in cases of individual self-deception and in cases of institutional self-deception do not necessarily stand or fall together. I will also argue that the kinds of difficulties involved in attributing responsibility to the self-deceived do not apply in instances of lying to oneself, which are intentional in nature, and so I will argue that institutions which lie to themselves can more straightforwardly be held responsible for the outcomes of doing so.

Finally, while I am unable to provide tried-and-tested solutions to the problem of institutional self-deception and institutions lying to themselves, I will end this Chapter with some suggestions for possible preventative measures that institutions can put in place to help guard against such undesirable states in Section 3.

Section 1: Individual Responsibility and Institutional Responsibility

We saw in Chapters 3 and 4 that institutions are the bearers of beliefs and desires, and capable of reflecting on and interrogating these states. They are capable of following (and violating) moral norms, and act in the world alongside other moral agents. In other words, institutions can be the appropriate targets for reactive attitudes like praise or blame. Indeed, in Chapter 4 we saw that institutions are capable of morally blameworthy acts, like lying. Moreover, we already have a history of blaming institutions, and not just individuals, for wrongdoing. It was Ford Motor Company after all, and not any individual, that was charged with three counts of reckless homicide in 1978.

I have argued that we are justified in holding institutions morally responsible for their actions. Nevertheless, one might worry that if one blames an institution for a wrongful act, then individual members of that institution will be absolved of any and all wrongdoing that they may have carried out. This need not be the case. List and Pettit suggest that there are three kinds of responsibility an individual could bear in relation to a group, namely a) designer responsibility, b) member responsibility, and c) enactor responsibility. Individuals may bear designer responsibility if they are responsible for putting the mechanisms of the institutional structure in place. Yet List and Petit argue that the institution will inevitably take

on a life of its own and change over time and, moreover, the relevant mechanisms may not be implemented as they were designed to be implemented, etc. As a result, List and Pettit do not regard this kind of responsibility as overly weighty, although it would be possible to apply such responsibility to individuals. List and Pettit also see member responsibility as fairly minimal. For example, while being a member of an institution that acts unethically can have moral implications for individuals – they ought to feel shame, perhaps – mere membership may not be something that one blames individuals too harshly for.[65] Enactor responsibility is the responsibility that accrues in virtue of being a member of the institution who contributes directly to – or carries out in full – the blameworthy action of the institution. List and Pettit regard enactor responsibility as the kind of responsibility that we ought to be most concerned about when thinking about the degree to which we should hold the individual members of an institution responsible for the blameworthy actions attributable to the institution itself. For List and Pettit, an important consideration here is that enactors carry out the relevant action when they needn't have done so – in other words, they could have refused to participate in or carry out the unethical actions of the institution. For example, we can imagine a situation in which an accounting firm puts pressure on its employees to fudge the numbers of particular clients when it is beneficial for them to do so. While there would likely be consequences for failing to do as they were told, the employee does not have to contribute to this act. However, if the member of an institution genuinely could not refuse to do their part in a blameworthy institutional act, then the degree to which they will be deemed responsible will naturally be diminished (List & Pettit, 2011, p164).[66] Thus List and Pettit argue we are capable of attributing responsibility to individuals as well as institutions for institutional acts. By way of extreme example, if an institution decides to eliminate its competition by assassinating the leader of a rival company, then the individual who pulls the trigger will have to answer for the part that they played – but the institution itself will also bear responsibility for master-minding and carrying out a murder. For List and Pettit, "members have responsibility as enactors of the corporate deed so far as they could have refused to play that part and didn't.

[65] Although it seems to me that assessing responsibility on the basis of mere membership can become quite complicated by considering factors such as an individual's knowledge or ignorance of the wrongdoing, and the degree to which continued membership (if voluntary) hints at some kind of implicit sanctioning of, or indifference to, unethical behaviour.

[66] Exactly what circumstances count as diminishing responsibility here may be difficult to determine – we may for example think that if an employee depends upon her job to survive, or raise her family, then she had no choice but to do her bit at the institution in order to keep her job. On the other hand, it may be the case that an employee's career or reputation is at stake if they refuse to carry out some or other task. Some may think that these concerns do not leave the employee without other options, however unpleasant the sacrifice may be for the employee to bear.

The group agent as a whole has responsibility as the source of that deed, the 'planner' at its origin" (List & Pettit, 2011, p163). Thus, as Gilbert would have it, in wanting to hold groups like institutions responsible for wrongdoings, "we are not forced to deny that any of the individual members bears guilt" (Gilbert, 2000, p151).

Christopher Kutz also discusses to what degree we can hold individual members of an institution responsible for institutional wrongdoing. Like List and Pettit, Kutz argues that when one can identify particular individuals within an institution whose individual actions contributed to the harm in question, then it would be quite reasonable for these individual members to bear responsibility. Factors which could come into play when assessing the distribution of responsibility within an institution, or when considering holding individual members of an institution responsible for some or other institutional wrongdoing, could include "relative ability to pay, voluntariness of the association with the organisation, fault, role-bound duties, and causal contribution" (Kutz, 2000, p201). If a company is fined for some wrongdoing, then those who earn the most at the company (ordinarily those who hold decision-making positions) could be made to bear the brunt of the fine (via pay cuts, etc.), while members who did not contribute to the harm could conceivably be (relatively)[67] unaffected by the institutions need to pay the fine. Alternatively, one can hold certain sectors of an institution responsible for some or other wrongdoing, depending upon the relevance of the duties they carried out to the harm. In addition to Ford having to compensate victims, for example, one might think that the members of the team who ran the safety tests on the Ford Pinto ought to be fired (or jailed) if it turned out that they deliberately withheld test results that indicated that the Pinto was unsafe. It therefore needn't be the case that attributing responsibility to institutions absolves individuals of responsibility for their actions and the role they play in institutional acts.

Kutz claims that in some cases, even though members have not personally contributed to the harmful act, they may still be punished insofar as the institution is punished. For example, if an institution needs to compensate a victim, everyone in the institution could be expected to take a pay cut so that the institution can afford to do so, regardless of the actual contributions of particular members to the harmful act. While this is perhaps a less desirable outcome than the former situations outlined above, if one did not hold the institution responsible here one would run the risk of a 'responsibility deficit' (Pettit, 2007; List &

[67] Although as List and Pettit suggest, they might feel guilty about their membership, and perhaps may still suffer reputational damage in terms of being associated with an unethical institution.

Pettit, 2011, p166) – a concern which was raised in Chapter 3. That individuals can bear the cost of the unethical behaviour of their institution may have the positive effect of making members of institutions particularly interested in ensuring that their institutions behave ethically. There is another benefit to be had from holding institutions themselves responsible for wrongdoing, namely minimising the possibility of 'corporate scapegoating' (Wilson, 1993). There are arguably situations where institutions have managed to avoid proper sanction by laying the blame at the feet of an individual. One such example of this is the case of the Exxon Valdez oil spill discussed by Wilson. When the oil tanker named Exxon Valdez ran aground in Alaska in 1989, it was the captain of the oil tanker, Joseph Hazelwood, who was blamed for the incident (Wilson, 1993, p784). He was purportedly known to be an alcoholic, and blamed for the accident on that front. The fact that the company had placed "an alcoholic captain in command of an oil tanker" (Davidson, 1990, p75) among other things, was swept under the rug. Wilson warns against a strong inclination toward methodological individualism, where explanations are given in terms of individuals rather than groups, as it may inhibit our ability to carry out "accurate responsibility ascriptions" (Wilson, 1993, p779). It might turn out upon closer inspection that the institution ought to bear the brunt of the blame. Indeed, as Wilson notes, "what was rarely asked was whether the company along with an entire industry was guilty of gross negligence. If that were so, then the captain may simply have been in the wrong place when there was an impending disaster" (Wilson, 1993, p784). There are parallels to be made with Pettit's concerns regarding the dearth of responsibility in the case of the Herald Free Enterprise, where no individual was able to be found properly responsible for the systematic negligence that led to the sinking of the ferry and the death of 193 people. If we only see individuals as morally or legally responsible for wrongdoing, then institutions can dodge responsibility, perhaps pay out a bit of money to compensate victims, but not have to make any changes to their behaviour or the processes that lead to such tragedies.

Lawford-Smith has also noted that being able to hold certain groups responsible for wrongdoings can be useful because "in some cases individual contributions may not themselves be wrongs, or harms, and so may only be punishable with a story about collective punishment" (Lawford-Smith 2018, p329).[68] There are also situations where a number of individuals have committed what could be regarded as minor wrongs, and so they can still be

[68] Kutz makes a similar point when he notes that "the culpability of an organization may also be greater than the culpability of its members" (2000, p198).

punished for their individual actions. However, Lawford-Smith argues that it will sometimes be the case that when their individual acts are taken together, one can see that a greater wrong has been committed – and thus, in being able to talk of group, or institutional wrongdoings, one can offer an explanation or a story which is "not just different, but better, in at least some cases" (Lawford-Smith, 2018, p329).[69]

It is also the case that if we can hold institutions to account, we may see solutions to problems that we may not have considered before. For example, instead of punishing or getting rid of a few (individual) bad apples, one may be more open to fining the actual corporation, or forcing a restructuring of the institution. One may decide that certain laws or regulations need to be amended in order to ensure changes in institutional behaviour. One may even be more open to dissolving problematic and recalcitrant institutions. While Lawford-Smith (2018) has argued convincingly for the permissibility of collective punishment, she is nevertheless cautious given the practical complications that arise in determining how to punish institutions. Indeed, accurate attributions of responsibility to institutions versus individuals is unlikely to be straightforward or obvious, and so would require a great deal of careful consideration. Nevertheless, she has noted that it is theoretically possible to do so, and given the power that many institutions wield, collective punishment may be a useful tool to help keep institutions in check.

Even if it is appropriate to hold institutions responsible for wrongdoing, there are nevertheless circumstances in which we do not think that agents are the appropriate targets of our reactive attitudes. Peter Strawson coined the term "reactive attitudes" (1962, p6) to refer to attitudes, such as anger or gratitude, of "offended parties and beneficiaries" (Strawson, 1962, p5), which he argues underpin our practice of holding people morally responsible. Yet Strawson suggests that there are certain considerations to take into account, like ignorance, coercion, incapacity, insanity, "or other less extreme forms of psychological disorder" (Strawson, 1962, p3) which can all obstruct attributions of moral responsibility. Thus, while an institution lying to itself may not be able to get off the hook (see below), one may worry that when it comes to self-deception in particular, responsibility attributions may not be

[69] She cites the example of a young black man who was assaulted and killed by a group of white youths in a racially motivated attack, and suggests that when the actions of the white youths is looked at on an individual basis one reaches the verdict of "assault with a deadly weapon, racist hate speech, and accessory to murder" (Lawford-Smith, 2018, p328-9), yet when one sees the group of white youths as a collective agent (while acknowledging that it is debatable whether or not it is justified to do so) one reaches a verdict of murder, and "Murder is more wrongful than assault with a deadly weapon, so collective punishment gets us more punishment for wrongs, rather than forcing us to take what we can get on an individualist analysis" (Lawford-Smith, 2018, p329).

appropriate – i.e., could self-deception be one of the 'less extreme forms of psychological disorder' which Strawson suggests would undermine moral responsibility? If this is the case then, even more worryingly, in arguing for the possibility of institutional self-deception I may have just handed over the best possible excuse to any institution which behaves unethically: 'we were self-deceived.' In the following section I hope to show that this is not the case.

Section 2: Lying to Oneself, Self-Deception and Institutional Responsibility

Self-deception is generally regarded as an undesirable state that can blind one to one's own moral shortcomings, and lead to problematic, harmful behaviour.[70] Yet the appropriateness of attributing moral responsibility to those who succumb to self-deception – either for their self-deceptive state or for the harm that self-deception may lead to – is a topic of some dispute. Levy (2004), for example, argues that attributing moral responsibility to the self-deceived only makes sense on a traditional (or literal), and now largely abandoned, understanding of self-deception as something intentional – the kind of model that I have referred to as lying to oneself. For Levy, one bears responsibility for intentional actions, and thus an institution that lied to itself could be regarded as responsible to some degree for their state and resulting actions because lying to oneself is something deliberate. It is not that straightforward, however. In fact, Levy observes that even views of self-deception in which the agent is taken to be split into different selves or parts (where one self, or part, is the deceiver and the other self, or part, is the deceived), it is common for the "deceived system [to be] identified with consciousness; with the agent's real self" (Levy, 2004, p302). In other words, even when considering quite literal takes on self-deception, or what I have been referring to as 'lying to oneself', one can view the agent in question as lacking in responsibility because the genuine agent is identified with the part of themselves that has been taken in by the lie rather than with the part of themselves that lied – in other words, they are the victim, not the perpetrator.

As discussed in Chapter 5, when it comes to individuals these intentional and literal accounts of self-deception are problematic for a number of reasons. However, given that institutions can lie to themselves, that one could then apportion responsibility within the institution accordingly and regard the institution as a victim of one of its parts may be

[70] However, I have alluded to the fact that there is also a trend, particularly in psychological literature, which regards self-deception as something beneficial – even essential – to human well-being (see Taylor 1989; Brown & Taylor, 1994; Kirsch, 2005). It is possible that the same could be said for institutions, but I do not have the space to discuss this possibility here.

reasonable. Such a proposal will inevitably become complicated by details regarding which parts of the institution did the deceiving and which did not, how essential these parts are to the identity of the institution, or the importance they play in setting the institutions agenda and orchestrating its actions. If those higher up the decision-making hierarchy are the one's doing the deceiving, the institution will look less like the victim than the perpetrator. Yet one could arguably regard the parts of the institution that were deceived as innocent victims, and the parts that did the deceiving as culpable. I suggested above that if the team in charge of conducting safety tests on the Ford Pinto had deliberately concealed the test results that indicated that the Pinto was in fact unsafe, or put forward false information, one may think that they should lose their jobs, or even face imprisonment. It is because of their actions that Ford came to hold the view it did regarding the safety of the Pinto. If the test results were suppressed by Ford's decision-makers one might think the same, or, depending on the details, one might be more inclined to identify their actions with those of Ford. In both situations it is possible that one can fall short of blaming Ford outright for releasing a vehicle it knew was unsafe (the suggestion being that Ford did not know, only a part of Ford knew).

Yet even in the best-case scenarios for Ford – where Ford is not identified with either of the potential perpetrator subgroups – Ford itself can still bear the blame for providing a working environment in which this sort of thing could take place. Given that it was Ford itself, and not just the safety unit, that decided to release the dangerous Ford Pinto, it is not clear that the team in charge of conducting safety tests should bear all the blame for the release of the car and the subsequent harm this caused. A department of an institution only has the power or influence that it does in virtue of its roles, responsibilities and capacities within that institution. But an institution can limit this power and influence, or impose checks and balances, and even get rid of a department. That deliberately withholding information, or providing false information, or suppressing information is possible within a given institution is reason for the institution itself to come under scrutiny. It may turn out that institutional structure or policy needs to changed. Perhaps more oversight built in. Consequently, when it comes to lying to oneself, I argue that we can rightly regard institutions as bearing some measure of responsibility.[71]

Deflationary accounts of self-deception, on the other hand, do away with the notion that deception is intentionally brought about, and so could appear to do away any degree of

[71] Of course, it may also turn out that upon closer examination that what at first appears to be an institutional wrongdoing could actually be an individual wrongdoing.

responsibility whatsoever. There has however been resistance to the argument that we are unable to attribute moral responsibility to the self-deceived if we adopt a deflationary account of self-deception.

Mele, for example, has argued for retaining responsibility by saying that it is possible for individuals to combat self-deception – a failure to do so would then make one blameworthy. The same can hold true for institutions. While we may enter or sustain self-deception via motivational evidence gathering, Mele notes that we are not wholly at the mercy of our desires and motivations when it comes to belief formation (Mele, 2001, p102-3). Individuals, and institutions, can guard against things like confirmation bias (one of the psychological phenomena that can contribute to self-deception) by instructing themselves to pay attention to both the pros and the cons of the matter at hand, or even just making themselves think about whether or not there are any grounds for doubt about whatever is being considered. This is especially true when one recognises that the matter being considered is important, and one realises that one is invested in a particular outcome being the case (Mele, 2001, p103). Ford, or the team carrying out the safety tests, should have realised that the pressure to release the Pinto as soon as possible would make them more invested in a particular outcome (that the car was safe). Consequently, they ought to have been careful to assess whether or not there were any reasons to doubt that the test results that indicated the car was safe were accurate and reliable. If they had done so, perhaps they would have paid more attention to the test results which indicated that the car was in fact not sufficiently safe. In addition, if an agent thinks that their motivations may be blinding them to the facts, they can check their views against those of others who are less invested in the matter at hand (Mele, 2001, p103). While it might seem tricky for companies to have members who are not as invested in a project or a particular outcome as others, Ford could have had someone who was not part of the Pinto project look over all the test results. If Ford created the safety team in such a way that biased reporting was possible, or if the safety team failed to recognise the import of certain data because of corporate structure and policy, then it is possible to blame Ford for its self-deception and the harms that resulted from it.

Yet Levy has argued that expecting an agent to guard against self-deception in the ways Mele suggests is too demanding. He contests that self-deception is only successful because the agent is unaware that their mental states – certain desires, for example – are making them more inclined to believe p than $\sim p$. If someone is aware that a desire they have is biasing their belief formation process, then they are more likely to be careful, and less likely to succumb to self-deception. On the other hand, if someone is unaware that they are

not, say, accurately weighing up the evidence – in fact, they very much think they are – then it is difficult to see how they can take steps to address their shortcomings in this regard. Consequently, Levy does not think we will typically be able to hold people responsible for their self-deceptive beliefs, as "responsibility for self deception requires that we take steps to reduce the grip of biasing mechanisms on our thought, but we can do so only when we are, or might become, aware of their presence" (Levy, 2004, p309). Furthermore, and according to Levy, when we are aware of their presence, we are unlikely to have fallen prey to self-deception.

Like Levy, Anna Galeotti is not convinced that you can satisfactorily make an agent responsible for their epistemic states which lead to self-deception when these states are not fully under the agent's control (Galeotti, 2016, p421). She echoes Levy's concerns with Mele's suggestion that if we know about certain cognitive biases, and are capable of vigilance, we should be held responsible when our vigilance fails. Galeotti adds to Levy's criticism by pointing out that there is nothing obviously wrong or irrational with the desires that tend to set off self-deception. For example, it is not unreasonable or irrational for Ford to desire that the Pinto be released as soon as possible in order to ensure its success in the market (to take advantage of the fact that consumers are currently interested in smaller vehicles – especially when this could change - and to beat its competitors etc.). Ford needs to be successful in order to continue to exist. As a result, it may in fact be extremely difficult to guard against self-deception.

One might respond that mere awareness of one's desires ought to be enough to trigger some extra vigilance – for example, Ford's being aware of its desire for the Pinto to be released as soon as possible should lead to extra vigilance in all matters relating to getting the car released on time. Even though it is not obvious that there is anything wrong with having such a desire, it could be argued that being aware of the desire should make Ford extra cautious in all matters regarding the release of the Pinto, and not, potentially, rushed and sloppy. Yet Galeotti is critical of the expectation that agents should know which desire or mental state they need to be on guard against – and being permanently in a state of hyper-vigilance seems a somewhat unreasonable a demand. That a great many of our desires or other mental states can go on to adversely affect cognitive processes in such a way as to give rise to self-deception is certainly a problem, but it is unclear that one could or should be expected to actually identify and guard against every single desire or mental state that could trigger self-deception. Furthermore, as Galeotti points out, it seems inappropriate to blame someone for failing to guard against a very normal emotional response to a situation.

While this might be a fair observation when it comes to individuals, it is harder for an institution to get off the hook this way. It is true that if Ford was aware that its desire to release the Pinto as soon as possible might have a biasing effect on its belief formation, then it would be more likely to be careful, and less likely to succumb to self-deception. It may also be the case that if Ford, or the team conducting the safety tests, is unaware that they are not accurately weighing up the evidence – indeed, they very much think they are – then they may struggle to take steps to address this issue. It even seems reasonable to suppose that Ford could not be expected to guard against a very rational desire. Yet as suggested above, it is possible for Ford to engineer a situation such that the team conducting the safety tests has to pass the results on to someone in the company who is less involved in the project, and thus more objective. Ford should be structurally organised in such a way as to limit the possibility of biased reporting by certain departments or teams. That Ford did not take steps to guard against the possibility of self-deception in terms of its structure and policies makes Ford responsible, at least to some degree, for self-deception here.

It has been suggested that agents can bear responsibility for being the kind of person who is easily self-deceived – i.e., for not having done enough to rid themselves of the sorts of character flaws that make one particularly or frequently susceptible to self-deception (Levy, 2004, p309). Levy is again critical of the suggestion that individuals can be responsible for being the kind of person who succumbs to self-deception. He is sympathetic to the view that deploying a certain coping mechanism, like self-deception, may well become reinforced in our characters over time with each successful instance of its deployment. However, as he points out, when it comes to individual instances of self-deception, it is very hard to see how we could be responsible for them. And if we are not responsible for individual instances of self-deception, then it becomes difficult to see how we can be responsible for a character which would seem to result from succumbing to these individual instances of self-deception. Yet when it comes to institutions, there is an argument to be made that they can bear responsibility for their characters. Institutions are artificial entities, after all. Thus, if the rules and regulations built into the fabric of an institution did not allow for any checks on important decision-making, or if the institution's policy were that it did not tolerate criticism of a decision after it had been made, then one can argue that institutional character is at fault. The institution can be to blame here because such policies seem clearly problematic – for example, the institution would not be open to all available evidence or views on important matters, or it would not consider relevant information when making decisions if it came to

light after a certain point – and because, importantly, the institution need not have the policies that it does.[72]

Dana Kay Nelkin provides further reasons to think that the self-deceived can be morally responsible, even without intentionality. She argues that we sometimes think that individuals can rightly be held responsible for harms that they cause, even when done unintentionally. For example, in situations where an agent may not have intended to cause harm, but ended up causing harm anyway because they acted unthinkingly, or behaved recklessly, they can rightly be blamed if we feel that they *should* have been aware of the (potential) harm their behaviour could cause. With this in mind, if one can determine whether or not a self-deceived individual or self-deceived institution should have been aware of the harm their action would cause, then one may be able to hold them responsible in those situations (Nelkin, 2012, p127-8). Thus, Nelkin notes that "if the criterion for responsibility is not awareness, but rather that one 'should have known', and if self-deception at least sometimes satisfies this criterion, then it is possible to see self-deception as a case in which people are sometimes responsible." (Nelkin, 2012, p127). However, Nelkin thinks it will be a complicated matter to determine when it is in the fact the case that one can reasonably say of the self-deceived that they ought to have known better, and that one would need to judge such matters on a case-by-case basis to determine whether or not responsibility can rightly be attributed to the self-deceived in any particular instance. With regard to the release of the Ford Pinto then, one can hold Ford responsible for their self-deceptive belief that the Pinto was a safe vehicle if one thinks that they should have known that it was not. Given that the test results which indicated the Pinto was not safe existed, one can argue that Ford should have known of them. If this is the case then, for Nelkin, self-deceived Ford can rightly be blamed for its conduct. However, it may be difficult to determine this with regard to many other institutions in similar situations that we are inclined to blame for harmful acts. When it comes to determining whether or not institutions ought to have known better one not only has to take into consideration an institution's meeting, or failing to meet, certain epistemic standards, one would also need to practically investigate what information the institution has on record in various departments, and perhaps form an assessment of its communication systems. Here, even if turns out that an institution could not have been aware of certain test

[72] Although it is possible that responsibility here could, however, lie with individuals rather than the institution, as highlighted by List and Pettit's discussion of 'designer responsibility'. However, given that institutions are usually able to review their policies and make changes to their procedures, there is room for institutional responsibility for institutional character.

results given the reporting structure in place, the institution could still be responsible for self-deception insofar as it is responsible for having a structure that lends itself so readily to self-deception. Nevertheless, it can still be said that for cases in which we do determine that institutions ought to have known better, self-deception will not be something which can excuse bad behaviour.

It is also possible to argue that institutions may have more of a responsibility than individuals to know better, and not to be self-deceived. Being competent with regard to their activities is important – and not just because of the number of people it affects when institutions get it wrong. Institutions also often occupy a unique epistemic position when it comes to their own activities – no one but Ford could realistically know about the safety of the Pinto leading up to its release, and no one but the Catholic church can know about how common child abuse is amongst priests, or how it is responded to. Consequently, institutions may have more of a duty to get things right than individuals do.

Given that there may still be some difficulty in determining whether or not an individual or an institution really ought to have known otherwise, and given that Nelkin seems to think that responsibility hinges on just this, I would now like to turn to what I take to be one of the most interesting and convincing accounts of responsibility for self-deception which is put forward by Anna Galeotti. Like Nelkin, Galeotti thinks that is important to retain some sense of responsibility for self-deception – after all, self-deception can have serious and harmful consequences (Galeotti, 2016, p420).[73] She argues that we can have responsibility for self-deception without it being the case that the agent is in control of the self-deception itself, instead viewing self-deception as an unintentional outcome of an agent's actions. And like Nelkin, she notes that we don't only apportion responsibility for outcomes that occur after rational deliberation, but also to those that result from inattentiveness, or carelessness, or thoughtlessness or impulsiveness (Galeotti, 2016, p426). Thus Galeotti argues that "responsibility can be attributed when someone performs a wrong or foolish act not consciously in the following conditions: (a) the agent has evidence that the act is wrong or foolish at the time; (b) she fails to recognize that evidence; (c) the failure falls below some applicable standards; (d) the faulty act is imputable to the agent." (Galeotti, 2016, 430). With the above in mind, Galeotti thinks we can hold onto moral responsibility for self-deception, and I am inclined to agree. If an agent has a self-deceptive belief that *p*, they will have access to evidence that suggests that this belief is false, that *~p* is in fact the case, yet they will fail to

[73] As the previous chapters have highlighted.

recognise that this is what the evidence suggests at the time. Impartial observers will be able to recognise that the evidence suggests that ~*p*, and thus it would seem that the failure of the agent to recognise this entails falling below certain standards we have regarding belief-formation. This accords with Mele's own recommendation that one employ an 'impartial observer test' as it requires one to determine whether or not the failing that led to the self-deceptive belief is a failing on the part of the particular agent, or whether it is an error anyone would make (which would imply that their cognising isn't [obviously] subpar).[74] And finally, for Galeotti, that the agent believes *p* and not ~*p* is appropriately attributable to the agent – i.e., is not the result of some external or coercive force. Here Galeotti explains that "whenever the agent fails to grasp the evidence for seeing his action as wrong or foolish and the failure can be attributed to him and not to some external or contingent impairments, then he can and ought to be held responsible for his action" (Galeotti, 2016, p427). Galeotti is suggesting that if an agent can rightly be deemed the originator of an action, and there wasn't some external force such as coercion or brainwashing at play, then the agent can be held responsible for their action – no matter if that action was performed out of ignorance or foolishness.

To elaborate, an agent like Ford will only not be responsible for an action like missing an international trade meeting if they missed it because of some external factor – for example, if a flight was cancelled due to volcanic eruptions and their representatives were thus unable to be at the meeting. To help determine whether or not the fault is caused by an external factor, Galeotti suggests that if an agent can look back on their misdeed (missing the meeting), reflect on their behaviour and determine whether there is a possibility that they can genuinely endeavour to do better in the future, then they can rightly be held responsible for said misdeed. Thus if they didn't miss the meeting because their flight was cancelled due to volcanic eruptions, but rather because they forgot to book a flight in good time, they can be held responsible because they can reflect on what they did and endeavour to do better in the future. Thus, for Galeotti, we can hold agents responsible if it is possible for them to learn from such misdeeds. Conversely, Galeotti says that "[i]f there is nothing to learn for the subject, I doubt that responsibility can be assigned" (Galeotti, 2016, p430). It is hard to see what could be learned from reflecting on what led to the misdeed (missing the appointment)

[74] Mele suggests that we employ the 'impartial observer test' when determining whether or not some case is an example of self-deception (2001, p106-7). If an impartial observer (someone who is a cognitive peer of the subject, but importantly lacks the relevant desire that p, or that ~p) comes to hold the same belief then the subject will not be considered self-deceived, but if, lacking the relevant desire they do not come to hold the same belief, then the subject will be considered self-deceived.

if this was due to external factors, as these factors tend to be beyond the control of the agent. Given this, attributions of responsibility to self-deceived institutions can be justified.

Section 3: Potential Solutions

Encouragingly, Galeotti notes that, unlike lying, or honest mistakes, self-deception is in principle easier to prevent because, while lying and mistakes aren't straightforwardly predictable, self-deception tends to be set off by a particular set of circumstances. If one can identify those circumstances and realise when those conditions are in place, one can be on the lookout for any potential self-deception (Galeotti, 2016, p431). Her approach differs somewhat from Mele's insofar as she recommends employing certain tools to combat self-deception in addition to being vigilant of desires and other mental states that tend to give rise to self-deception. To combat self-deception, Galeotti suggests a combination of character building and precommitment. Character building would involve cultivating certain virtues like attentiveness, objectivity, humility and discipline etc. An institution that prized attention to detail and honesty could arguably fare better in terms of avoiding self-deception than one which prized loyalty, competitiveness, or pleasing one's customer base above all else. If those employees involved in performing the safety tests on the Ford Pinto had been impressed by the need to pay attention to detail, and been humble in drawing conclusions about the safety of the vehicle, then perhaps they would have been more likely to notice the worrying test results, or less likely to reach the conclusion that they did regarding the Pinto's safety. An institution with core values like honesty and integrity that permeated the workplace or institutional environment would presumably also be less likely to lie to itself. If honesty were highly valued and even rewarded at Ford, then the team that conducted safety tests on the Pinto would surely be less likely to deliberately withhold negative test results. Thus character building could also be of immense value when it comes to combatting the problem of an institution lying to itself.

There are a number of more practical measures one can take to encourage honesty in general in institutional settings, as outlined by Dan Ariely in his book, *The (Honest) Truth about Dishonesty: How We Lie to Everyone - Especially Ourselves*. A great number of the devices he recommends can also be of some benefit when it comes to curbing institutional self-deception (insofar as they would aid the building of an honest institutional character), but more pointedly, such devices could help to discourage an institution from lying to itself. Ariely's research and experiments have shown that reminders of ethical codes – for example,

being asked to recall the ten commandments, or to sign an honour code before completing a task – can help to curb unethical behaviour, at least temporarily. He found, however, that being asked to sign honour codes and being given crash courses in ethics did not have a long-lasting effect on behaviour (Ariely, 2012, p40-43).[75] As such, institutions could consider requesting employees to sign a declaration that they will abide by the institution's honour code (or some equivalent) when submitting key information, or when making certain important decisions, as doing so immediately prior to performing some or other task seems to effectively inhibit dishonesty. As Ariely notes, "when we are removed from any benchmarks of ethical thought, we tend to stray into dishonesty. But if we are reminded of morality at the moment we are tempted, then we are much more likely to be honest" (Ariely, 2008, p213). When it comes to situations in which an institution lies to itself, doing as Ariely suggests could be a useful preventative measure to put in place. If the team that conducted safety tests on the Ford Pinto is tempted to deliberately withhold negative test results because of pressure to release the car as soon as possible, having to sign a declaration that they have abided by the institution's honour code when submitting the final test results or their report on the safety of the Pinto may make it more likely that the team will rethink submitting a false report.

Ariely also discusses the importance of clarity when it comes to rules or a code of conduct – if expectations are vague, and rules are ambiguously phrased, institutions give their members a lot of leeway to rationalise their unethical behaviour (Ariely, 2012, p208; p219). Thus if the safety team at Ford is told to do its best, it could be up for debate whether or not doing their best means producing a good result (i.e., the result that Ford wants, that the Pinto is safe and can be released), or whether it means carrying out its duties with the utmost care and scrutiny (and so take longer to give the greenlight regarding the release of the Pinto). Ariely has even discovered that a surprisingly simple thing could improve institutional honesty – "the mere feeling of being watched can inhibit bad behaviour" (Ariely, 2012, p224), and thus something as simple as sticking up pictures of eyes in the workplace can encourage honest behaviour. As bizarre as it may sound, if fake eyes were posted in the testing zone, or at the desk where the safety reports were being filled out, it may have been

[75] In experiments conducted where participants are tempted to cheat on a test, no participant cheated when asked to recall the ten commandments or their institution's honour code prior to taking the test. However, many participants did cheat when asked to recall something neutral like their ten favourite books, just as they had in control sessions when not asked to do anything before writing the test (Ariely, 2012). Yet if participants took the test after completing an ethics course two weeks beforehand, they still cheated just as much as those in the control sessions, indicating that business ethics courses conducted annually (or biannually) are unlikely to make any long-lasting impact on behaviour (2012, p43-44).

more difficult for the team to deliberately withhold the negative test results and pass on a false report.

Worryingly, it can also be quite easy to cultivate the wrong kind of character. Ariely suggests that there is an "infectious element to dishonesty" (Ariely, 2012, p192) – if a particular individual within an institution bends the rules and behaves unethically, say, then over time other individuals exposed to their unethical behaviour are more likely to follow suit because they come to view such behaviour as acceptable – particularly if the unethical individual a) does not appear to face any repercussions for their actions (as long their actions were not *too* bad), and b) is someone employees identify with (Ariely, 2012, p199-204).[76] Thus unethical behaviour, like lying, cheating or even stealing can become a social norm within an institution (Ariely, 2012, p197), and contribute to the cultivation of a dishonest character – a character that for Galeotti may be more likely to succumb to self-deception, or I would argue, lying to itself. Consequently, institutions ought to take care not to tolerate or excuse bad behaviour, or they will run the risk of this becoming a defining aspect of the institution's character. As Ariely warns, "it's important to realize that the effects of individual transgressions can go beyond a singular dishonest act" (Ariely, 2012, p214).

Yet there are perhaps things we could do which focus more specifically on tackling the problems of institutional self-deception. Glaeotti for one does not see character building as sufficient to prevent self-deception, as even someone of good epistemic character may fall prey to self-deception in the right circumstances. Indeed, for Galeotti, the self-deceived isn't necessarily someone who is weak-willed or greedy or unable to bear reality etc., but rather "someone experiencing a lapse of self-discipline under special circumstances" (Galeotti, 2016, p432).[77] Moreover, as institutional self-deception need not be brought about by individuals doing wrong, generating a more virtuous workplace will only go so far toward avoiding it. Galeotti recommends precommitment in order to guard against such temporary lapses, having the agent take some measures beforehand to limit their likelihood of falling prey to self-deception in the future. She explains that in ordinary cases of self-deception peers of the self-deceived can often inadvertently end up supporting their self-deceptive beliefs. For

[76] Ariely for example discovered in a series of experiments that if participants expected to complete a task were exposed to someone who obviously cheated at the task, their own dishonesty would increase – unless the cheater cheated by some unacceptable/unjustifiably large amount, at which point, interestingly, they tended to react by being more honest. In addition, if the cheater was regarded as a member of the participants own social group then cheating increased across participants, yet if the cheater was regarded as an outsider (they wore a shirt from a different university to that which the participants belonged to, say), this served to inhibit dishonesty.

[77] When one considers how common self-deception appears to be (Taylor, 1989) this lends support the view that self-deception is not an affliction of just those of particularly poor character.

example, because they may want to avoid any awkwardness that could arise from disagreeing, many peers do not actually voice any disagreement, and so are often regarded as implicitly agreeing with the self-deceived (especially since the self-deceived can also be motivated to interpret any evidence available as favouring a particular belief – a lack of dissent could therefore be counted as support for their belief). Peers of the self-deceived may also fail to voice disagreement because they don't want to upset someone by challenging a self-deceptive belief, and because they feel sorry for the self-deceived. Galeotti suggests that one can ensure that one will not merely be humoured by one's peers in this sort way by requesting of certain others that they act as honest referees for one's beliefs. While one may not be able to ensure that everyone around one does not merely humour one, one can appeal to family and friends, say, to be honest with one and correct or challenge one if the need arises. She emphasises that it's important that the agent "directly and explicitly confers on her special friend(s) the authority of referee(s) over her potential [self-deception]" (Galeotti, 2016, p433) because this is the best way of ensuring that those friends feel able to be honest and so challenge the self-deceptive beliefs. In conferring the authority to do so on those friends oneself, one is employing a precommitment strategy to combat self-deception (Galeotti, 2016, p434).

Institutions could do something similar. It is possible that engendering an environment where honesty is valued could make it easier for the institution to tolerate challenges to its views or policies, and could foster much needed internal criticism and self-reflection. Yet just as happens in individual cases, temporary lapses of these sorts of practices could still result in self-deception. In terms of precommitment, as suggested above it may be possible for institutions to structure themselves in such a way that someone in the company who is less involved in the project, and thus more objective, is required to look over certain project reports. One may worry, however, that any specialised watchdog-type unit within an institution will be able to be sufficiently separated from the circumstances which give rise to self-deception to be able to successfully challenge the self-deceptive belief – if they were also a part of the institution, they would also be likely to be taken in by, or subject to desires or other pressures that triggered the self-deception, like the desire for a particular project to be a success so that the company doesn't go under and cease to exist). Yet institutions could get outsiders – critical consultants – to serve as sounding boards, giving them the authority to be brutally honest. While it may be difficult to find an outsider who is familiar enough with all the relevant details, and while certain institutions may be reluctant to grant an outsider access to information for security reasons, say, this nevertheless remains a powerful tool that

institutions could employ to guard against institutional self-deception. Galeotti herself also suggests that something like this this could be useful in an institutional setting, while also gesturing toward the fact that there may well be a number of difficulties in place in terms of successful implementation (Galeotti, 2016, p434). However, for certain institutions whose actions have a very obvious and direct effect on the lives of others (i.e., governments), one may wish to go a step further and insist as Bok does that "more than consultation with chosen peers is needed" (1978, p97). Indeed, there is always the risk that when one is responsible for choosing one's critical peer, one can choose someone whose interests one knows align with one's own. In certain situations, deliberate consultation and discourse with the wider public may be necessary in order to avoid unconscious biases, or shared assumptions and faulty reasoning (1978, p96).

In this chapter I hope to have shown that holding institutions rather than individuals responsible for institutional self-deception and an institution's lying to itself is not only possible, but powerful when it comes to combating certain social ills that institutions are so often at the heart of, insofar as holding institutions responsible for institutional wrongdoing is often necessary if we want to ensure that the wrongdoing does not go unpunished. Given the intentional nature of lying it is perhaps easier to see how an institution would bear responsibility for lying to itself – and any harms that result from doing so – however it is also the case that being self-deceived will not let institutions off the hook. Nevertheless, given the complexity of assessing what exactly goes on in institutional settings, it will remain a difficult task to accurately identify when either of these situations have occurred. It is, for example, still possible that what may appear to be an example of an institution lying to itself is in fact better understood as mere individual interpersonal deception, and an institution that is self-deceived and an institution that is simply lying to the public can have the same outward appearance in terms of their assertions.

In spite of these challenges, given the manifest undesirability of institutional self-deception and institutions lying to themselves, it is somewhat encouraging that there may well be some preventative measures that we can take. That we can prevent or inhibit institutional self-deception and an institution lying to itself from taking place is particularly useful given the difficulty of accurately identifying each of these phenomena after the fact – when the stakes are high, it may be that if we adopt the stratagems suggested by authors like Ariely, and take Galleotti's suggestion of an external watchdog seriously, then these are problematic scenarios that we may not need to worry about at all.

Conclusion

A great deal of recent philosophical work has argued for seeing certain groups as agents, capable of beliefs and desires, and as able to make assertions and to lie. Recently it has also been suggested that self-deception is something that affects groups as well as individuals. Taking such suggestions seriously, my aim has been to determine whether or not a certain category of group that I regard as a particularly promising candidate for mindedness and agency – institutions – are capable of being self-deceived, or lying to themselves.

As we have seen, however, self-deception is a topic that is fraught with controversy – in part because self-deception has traditionally been modelled on interpersonal deception, where the self is both the deceiver and the deceived. The problems that arise from this puzzling situation are done away with with the introduction of deflationary accounts of self-deception such as Alfred Mele's, which see self-deception as involving a false belief that arises from unintentional motivationally biased evidence gathering. One of the most obvious ways, then, of determining whether or not institutions could be self-deceived was to see whether or not institutions are capable of meeting Mele's four minimal conditions for self-deception. I have argued that they can.

While I noted that the suggestion that institutions could be self-deceived has put forward by a number of theorists looking at the unethical behaviour of certain institutional bodies, the discussions of this phenomenon seems primarily to have regarded it as something which could be reduced either to talk of the head of an institution being self-deceived, or to talk of individuals in an institution that are similarly self-deceived about some or other proposition, or to cases of interpersonal deception within an institution. In arguing that an institution is capable of beliefs of its own, and thus that it is also capable of self-deceptive beliefs of its own, I hope to have provided an alternative account.

Yet I have also argued that because institutions are made up of a number of individuals or sectors, and tend to be organised hierarchically with a differentiation of roles and responsibilities, institutions are also capable of instantiating what could be deemed a more traditional understanding of self-deception – i.e., they can quite literally lie to themselves. Having these additional explanations in our philosophical toolkit could open up interesting questions surrounding government responses across the world to the Covid-19 pandemic, or the institutional spread of 'fake news', and offer alternative explanations for a given institution's denial of climate change concerns, particularly when there is a case to be made that the charge of being a liar does not seem to fit. Furthermore, while I established in

Chapter 1 that the focus of this thesis would be on the possibility of institutional self-deception, and an institution's ability to lie to itself, it may well be the case that one or the other of the above phenomena could afflict a much wider range of groups than I have considered here.

I have taken both institutional self-deception and an institution lying to itself to be at least potentially problematic phenomena, and something which could be behind some instances of problematic and unethical institutional actions. With this is mind, I also hope to have shown that holding institutions rather than individuals responsible for institutional self-deception and an institution's lying to itself is not only possible, but a potentially powerful practice. In addition to this, I have put forward a number of recommendations for preventing or remedying institutional self-deception and an institution lying to itself.

Having stressed the potential harm of institutional self-deception, it is nevertheless possible that not all self-deceptive beliefs are in fact harmful. There are suggestions that self-deception may actually be extremely beneficial – that harbouring self-deceptive beliefs can "contribute to the production of better mood, better popularity, better ability to care for others, creativity, productivity, resilience from stress, and ultimately happiness" (van Leeuwen, 2009, p107), albeit if kept within certain healthy limits. That this could be true of institutions as well as individuals, and could be a topic worth pursuing in future research. If true, this may serve to change and indeed complicate our attitude towards the discovery of institutional self-deception.

References:

Aßländer, Michael S. (2013) "Honorableness or Beneficialness? Cicero on Natural Law, Virtues, Glory, and (Corporate) Reputation", *Journal of Business Ethics* 116: 751-767.

Anderson, Nic. (2018, 15 February) "DA left embarrassed as Patricia De Lille survives motion of no confidence", *The South African.* Available at: https://www.thesouthafrican.com/news/patricia-de-lille-survives-motion-of-no-confidence/ (accessed: 2 January 2020).

Ariely, Dan. (2008) *Predictably Irrational: The Hidden Forces That Shape Our Decisions.* HarperCollins.

- (2012) *The Honest Truth About Dishonesty: How We Lie to Everyone – Especially Ourselves*. HarperCollins.

Armstrong, G., Gurol, M., & Russ, F. (1979). "Detecting and Correcting Deceptive Advertising", *Journal of Consumer Research* 6 (3): 237-246.

Becker, P.J., Jipson, A.J. and Bruce, A. (2002) "State of Indiana V. Ford Motor Company revisited", *American Journal of Criminal Justice* 26 (2): 181-202.

Block, Ned. (1978) "Troubles with Functionalism", *Minnesota Studies in the Philosophy of Science* 9: 261-325.

- (1990) "Inverted Earth", *Philosophical Perspectives* 4: 53-79.

Bok, Sissela. (1978) *Lying: Moral Choice in Public and Private Life*. New York: Vintage Books.

Bortolli, L. and Mamelli, M. (2012), "Self-Deception, Delusion and the Boundaries of Folk Psychology", *Humana.Mente* 20: 203-221.

Bratman, Michael. (2009) "Shared Agency" in C. Mantzavinos (Ed.), *Philosophy of the social sciences: Philosophical theory and scientific practice*. Cambridge University Press.

Chalmers, David & Clark, Andy. (1998) "The Extended Mind", *Analysis* 58 (1): 7-19.

Chisholm, Roderick M. and Feehan, Thomas D. (1977) "The Intent to Deceive", *The Journal of Philosophy* 74 (3): 143-159.

Cohen, LJ. (1989) "Belief and Acceptance", *Mind* 98 (391): 367-389.

Collins, Stephanie. (2019) *Group Duties: Their Existence and Their Implications for Individuals*. Oxford University Press.

Crane, Tim. (2010) "Tim Crane on Mind and Body" in Edmonds, D., & Warburton, N. *Philosophy Bites*. Oxford: Oxford University *Press.*

Davidson, Art. (1990) *In the Wake of the Exxon Valdez: The Devastating Impact of the Alaska Oil Spill*. Sierra Club Books.

Davidson, Donald. (1963) "Actions, Reasons, Causes", *Journal of Philosophy* 60: 685-99.

- (1970) "Mental Events" in L. Foster & J. W. Swanson (eds.), *Experience and Theory*. Humanities Press.
- (1982) "Paradoxes of Irrationality" reprinted in *The Essential Davidson*. (2006). Oxford University Press.
- (2001) "What Thought Requires" in Joao Branquinho (ed.), *The Foundations of Cognitive Science*. Oxford: Clarendon Press.

Davis, Lawrence. (1982) "What is it like to be an agent?", *Erkenntnis* 18: 195-213.

Demos, Raphael. (1960) "Lying to Oneself", *The Journal of Philosophy* 57 (18): 588-595.

Dennett, Daniel. (1971) "Intentional Systems", *The Journal of Philosophy* 68 (4): 87-106.

- (1981) "True Believers: The Intentional Strategy and Why it Works" reprinted in John Haugeland (ed.) *Mind Design II: Philosophy, Psychology, Artificial Intelligence* (1997). The MIT Press.

- (1988) “The Conditions of Personhood” in Goodman M.F. (ed.) *What Is a Person?* Humana Press.
- (1995) “Animal Consciousness: What Matters and Why”, *Social Research* 62 (3): 691-710.

Deweese-Boyd, Ian. (*forthcoming*), “Collective self-deception, collective injustice: Consumption, sustainability and responsibility”.

Dowie, Mark. (1977) “Pinto Madness” *Mother Jones* (2): 18-32.

Epstein, Brian. (2017) “What are social groups? Their metaphysics and how to classify them”, *Synthese* 196 (2): 1-34.

Finn, Kathy. (2013, 30 September) “BP lied about size of U.S. Gulf oil spill, lawyers tell trial”, *Reuters*. Available at: https://www.reuters.com/article/us-bp-trial/bp-lied-about-size-of-u-s-gulf-oil-spill-lawyers-tell-trial-idUSBRE98T13U20130930 (accessed: 2 February 2020).

Fisse, Brent and Braithwaite, John. (1987) “The Impact of Publicity on corporate Offenders: Ford Motor Company and the Pinto Papers” in Ermann, M.D. & Lundman, R.J. (eds.) *Corporate and Governmental Deviance*. Oxford University Press.

Frankel, Robin Saks. (2020) “Small Businesses Continue To Suffer Pandemic Pain. Is Bankruptcy The Answer?”, *Forbes Advisor*. Available at: https://www.forbes.com/advisor/personal-finance/small-businesses-continue-to-suffer-pandemic-pain-is-bankruptcy-the-answer/ (accessed: 20 July 2021)

Frankfurt, Harry. (1978) “The Problem of Action”, *American Philosophical Quarterly* 15 (2):157-162.

- (2004) *The Reasons of Love*. Princeton University Press.

Fraser, Douglas. (2019, 30 November) “General Election 2019: Do firms have purpose beyond profit?” *BBC News*. Available at: https://www.bbc.com/news/uk-scotland-50614999 (Accessed 10 January 2020).

French, Peter. (1979) “The Corporation as a Moral Person”, *American Philosophical Quarterly* 16 (3): 207-215.

Fricker, Miranda. (2012) “Group Testimony? The Making of A Collective Good Informant”, *Philosophy and Phenomenological Research* 84 (2): 249-276.

Galeotti, Anna. (2012) “Self Deception: Intentional Plan or Mental Event?”, *Humana.Mente* 5 (20): 41-66.

- (2016) “The Attribution of Responsibility to Self-Deceivers”, *Journal of Social Philosophy* 47 (4): 420-438.

Gilbert, Margaret. (1987) “Modelling Collective Belief”, *Synthese* 73: 185-204.

- (1989) *On Social Facts*. Princeton, New Jersey: Princeton University Press.
- (1990) “Walking Together: A Paradigmatic Social Phenomenon”, *Midwest Studies in Philosophy* XV: 1-14.
- (2000) *Sociality and Responsibility*. USA: Rowman & Littlefield Publishers Inc.

Gilbert, Margaret and Daniel Pilchman (2014). “Belief, Acceptance, and What Happens in Groups: Some Methodological Considerations”, in Jennifer Lackey (ed.) *Essays in Collective Epistemology*. Oxford: Oxford University Press.

Harris, Sam. (2011) *Lying*. Harvard Business School Press.

Haslanger, Sally. (2000) “Gender and Race: (What) Are They? (What) Do We Want Them To Be?” *Nous* 34, 1: 31-55.

Hill, Thomas. (1991). *Autonomy and Self-Respect*. Cambridge University Press.

Hoffman, W.M. (2014) “The Ford Pinto” in W. Michael Hoffman, Robert E. Frederick and Mark S. Schwartz (eds.) *Business Ethics: Readings and Cases in Corporate Morality.* Wiley Blackwell.

“institution, n.”. OED Online. December 2020. Oxford University Press. Available at:

https://-0www.oed.com.wam.seals.ac.za/view/Entry/97110?redirectedFrom=institution& (accessed January 5, 2021).
Jarry, Emmanuel. (2019, 21 January) "France holds trial of firms accused of selling horse meat as beef", *Reuters*. Available at: https://www.reuters.com/article/us-france-food-horse/france-holds-trial-of-firms-accused-of-selling-horse-meat-as-beef-idUSKCN1PF11V (accessed: 10 March 2020).
Kadlac, Adam. (2010) "Humanizing Personhood", *Ethical Theory and Moral Practice* 13 (4): 421-437.
Kaiser, Marie. (2015) *Reductive Explanation in the Biological Sciences.* Springer.
"KFC's apology for running out of chicken is pretty cheeky". (2018, 23 February) *BBC News*. Available at: https://www.bbc.com/news/newsbeat-43169625 (Accessed: 28 December 2019).
Kim, Jaegwon. (2000) *Mind in a Physical World*. The MIT Press.
- (2010) *The Philosophy of Mind 3rd Edition*. Routledge.
Kirsch, Julie. (2005) "What's So Great About Reality?" *Canadian Journal of Philosophy* 35 (3): 407-427.
Korsgaard, Christine. (1996) *The Sources of Normativity*. Cambridge University Press.
Kutz, Christopher. (2007) *Complicity: Ethics and law for a collective age*. Cambridge University Press.
Lackey, Jennifer. (2007) "Norms of Assertion", *Noûs* 41 (4): 594-626.
- (2018) "Group Lies" in Michaelson, Eliot and Stokke, Andreas (eds.) *Lying: Language, Knowledge, Ethics, and Politics*. Oxford: Oxford University Press.
Lawford-Smith, Holly. (2018) "What's Wrong with Collective Punishment?", *Proceedings of the Aristotelian Society* cxviii (3): 327-345.
Le Bon, Gustave. (2009) *The Crowd: A Study of the Popular Mind*. Mineola, New York: Dover Publications, Inc. (Original work published 1945).
Levi, Isaac. (1990) Hard Choices. Cambridge University Press.
Levy, Neil. (2004) "Self-Deception and Moral Responsibility", *Ratio (new series)* 17: 294-311.
Lelinwalla, Mark. (2016, 2 March) "Google Thinks Self-Driving Cars Will Be Ideal For Senior Citizens", *TechTimes*. Available at: https://www.techtimes.com/articles/138048/20160302/google-thinks-self-driving-cars-will-ideal-senior-citizens.htm (accessed: 2 January 2020).
List, Christian and Pettit, Philip. (2011) *Group Agency: The Possibility, Design, and Status of Corporate Agents.* Oxford University Press.
Ludwig, Kirk. (2014) "Proxy Agency in Collective Action", *Noûs* 48 (1): 75-105.
- (2016) *From Individual to Plural Agency: Collective Action: Volume 1*. Oxford University Press.
- (2017) *From Plural to Institutional Agency: Collective Action: Volume 2*. Oxford University Press.
Mele, Alfred. (2001) *Self-Deception Unmasked*. Princeton University Press.
- (2012) "When Are We Self-Deceived?" *Humana.Mente* 20:1-15.
Menkiti, Ifeanyi. (1984) "Persons and community in African traditional thought", in Wright, R. (ed.) *African Philosophy: An Introduction*. University press of America, Lanham, MD.
- (2004) "On the Normative Conception of a Person" in Wiredu, K. (ed)
Companion to African Philosophy. Oxford: Blackwell Publishing.
Miller, Norman. (2020, 13 April) "How factories change production to quickly fight

coronavirus", *BBC Worklife*. Available at: https://www.bbc.com/worklife/article/20200413-how-factories-change-production-to-quickly-fight-coronavirus (accessed: 29 June 2021).
Nagel, Thomas. (1974) "What is it like to be a bat?" *The Philosophical Review* 83 (4): 435-450.
Nelkin, Dana Kay. (2012) "Responsibility and Self-Deception: A Framework", *Humana.Mente* 20: 117-139.
Okasha, Samir. (2000) "Van Fraassen's Critique of Inference to the Best Explanation," *Stud. Hist. Phil. Sci.* 31 (4): 691-710.
Olson, Eric. (1995) "Human People or Human Animals?", *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 80 (2): 159 -181.
Pettit, Philip. (2004) "Groups with Minds of their Own", in Frederick Schmitt (ed.) *Socializing Metaphysics*. New York: Rowman and Littlefield.
- (2007) "Responsibility Incorporated", *Ethics* 117: 171-201.
- (2014) "How to Tell if a Group is an Agent" in Jennifer Lackey (ed.) Essays in Collective Epistemology. Oxford: Oxford University Press.
- (2014b) "Philip Pettit on Group Agency" in Edmonds, D., & Warburton, N. *Philosophy Bites Again*. Oxford: Oxford University Press.

Phakgadi, Pelane. (2018) "Tiger Brands Refuses to Take Blame for Listeriosis Deaths", *Eyewitness News*. Available at: https://ewn.co.za/2018/03/06/tiger-brands-refuse-to-take-blame-for-listeriosis-deaths (Accessed: 18 April 2018)
Piper, Adrian. (1992) "Passing for White, Passing for Black" *Transition* 58: 4-32.
Quinton, Anthony. (1975) "Social Objects", *Proceedings of the Aristotelian Society New Series* 76: 1-27
Ravenscroft, Ian. (2005) *Philosophy of Mind*. Oxford University Press.
Ritchie, Katherine. (2013) "What are Groups?" *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 166 (2): 257-272.
- (2015) "The Metaphysics of Social Groups", *Philosophy Compass* 10 (5): 310-21.

Roddick, Anita. (2000) *Business as Unusual: The Triumph of Anita Roddick and The Body Shop*. HarperCollins Publishers.
Rovane, Carol. (1994) "The Personal Stance", *Philosophical Topics* 22: 351-396.
- (2004) "What is an Agent?" *Synthese* 140: 181-198.

Rupert, Robert. (2014) "Against Group Cognitive States" in S. Chant, F. Hindriks, and G. Preyer (eds.) *From Individual to Collective Intentionality* Oxford: Oxford University Press.
Sagarin, Brad J., Rhoads Kelton V.L., and Cialdini, Robert B. (1998) "Deceiver's distrust: denigration as a consequence of undiscovered deception", *Personality & Social Psychology Bulletin* 24 (11): 1167.
Schaefer, Mark. (2018, 18 October) "Facebook has no moral compass". Available at: https://businessesgrow.com/2018/10/18/facebook-has-no-moral-compass/ (accessed 28 March 2019).
Schwitzgebel, Eric. (2015) "If materialism is true, the United States is probably conscious", *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 172 (7): 1697-1721.
Searle, John. (1980) "Minds, Brains, Programs", *The Behavioral and Brain Sciences* 3: 417-457.
- (2005) "What is an institution?" *Journal of Institutional Economics* 1 (1): 1–22.

Sharma, R. R., and Chander, S. (2011). "What's Wrong with Misleading Advertising? - An Empirical Investigation", *Asia Pacific Business Review* 7 (1): 191-205.
Shoemaker, Sydney. (1999). "Review of *The Human Animal*, by E. Olson", *Noûs* 33 (3):

496–504.
Shukman, David. (2018, 26 April) “Companies sign up to pledge to cut plastic pollution” *BBC News*. Available at: https://www.bbc.com/news/business-43901328 (accessed: 28 March 2019).
Sorensen, Roy. (2005) *A Brief History of the Paradox: Philosophy and the Labyrinths of the Mind*. Oxford University Press.
South Africa Government. (2021) *Register your company.* Available at: https://www.gov.za/services/register-business-or-organisation/register-your-company (accessed January 2, 2021).
Sprevak, Mark and Statham, David. (2015) “Group minds and explanatory simplicity”, *Royal Institute of Philosophy Supplement* 76: 3-19.
“Spy poisoning: Russia expels 60 US diplomats in tit-for-tat measure” (2018, 30 March). *BBC News*. Available at: https://www.bbc.com/news/world-europe-43590933 (Accessed: 18 April 2018).
Steinberg, Ted. (2007) “Mr. Corporation on the Couch”, *Environmental History* 12 (2): 372-374.
Strawson, Peter. (1962) “Freedom and Resentment” in *Freedom and Resentment and Other Essays*. Routledge (Reprinted in 2008).
Sweet, Matthew (Host). (2015) “Lying and Misleading” [Audio podcast episode], *The Philosopher’s Arms*. BBC Radio 4. Available at: https://www.bbc.co.uk/programmes/b06r4gpz.
Szabados, Béla. (1973) “Wishful Thinking and Self-Deception”, *Analysis* 33 (6): 201-205.
Taylor, Shelley (1989) *Positive Illusions: Creative Self-deception and the Healthy Mind*. Basic Books/Hachette Book Group.
Tollefsen, Deborah. (2015) *Groups as Agents*. Cambridge: Polity Press.
Trivers, Robert. 2000, “The Elements of a Scientific Theory of Self-Deception”, *Annals of the New York Academy of Sciences* 907: 114-131.
Turing, Alan. (1950) “Computing Machinery and Intelligence” *Mind* LIX 236: 433-460.
Van Leeuwen, Neil. (2009) “Self-Deception Won’t Make You Happy”, *Social Theory and Practice* 35 (1): 107-132.
Vinocur, Nicholas. (2013, 14 February) “French firm Spanghero says believed was selling beef, not horse”, *Reuters*. Available at: https://uk.reuters.com/article/uk-horsemeat-france-spanghero/french-firm-spanghero-says-believed-was-selling-beef-not-horse-idUKBRE91D19120130214 (accessed: 19 April 2020).
Wagner, Tilllman, Lutz, Richard J. and Weitz, Barton A. (2009) “Corporate Hypocrisy: Overcoming the Threat of Inconsistent Corporate Social Responsibility Perceptions” *Journal of Marketing* 73 (6): 77-91.
Weiner, Matthew (2005) “Must we know what we say?” *The Philosophical Review* 114 (2): 227-251.
Williams, Bernard. (2002). *Truth and Truthfulness*. Princeton University Press.
Williamson, Timothy. (1996) “Knowing and asserting”, *The Philosophical Review* 105 (4): 489-523.
Wilson, P. Eddy. (1993) “The Fiction of Corporate Scapegoating”, *Journal of Business Ethics* 12 (10): 779-784.