Introduction

Trust is important to living a flourishing human life. As Margaret Walker has pointed out, “The results of shattered trust can be lives in which people cannot sustain employment or relationships, or lose the ability to move about publicly without distress” (Walker, 2006, p. 93). Betrayal is significant to human flourishing insofar as it can damage trust. Trust’s vulnerability to betrayal has been identified as a feature which sets trust apart from mere reliance (Baier, 1986, p. 235; Hieronymi, 2008, p. 215; Holton, 1994, p. 66; McGeer, 2002, p. 33; McLeod, 2011; Walker, 2006, p. 85; Wright, 2009). However, there has been little direct philosophical analysis of betrayal or its implications for understanding trust. This lack of analysis leaves a gap in our understanding of the landscape of moral agency. A clear account of betrayal is needed to explain the distinction between trust and mere reliance, and to understand the damage that betrayal can inflict on trust.

One promising approach to understanding what betrayal is and how it operates is to consider instances of damaged trust. The shortfall of understanding betrayal in this way is that doing so requires a clear account of trust, which can in turn smuggle in assumptions about betrayal – for example, that betrayal distinguishes trust from mere reliance. Insofar as trust and betrayal are to some extent understood with reference to each other, teasing out the relationship between them is by no means straightforward. It seems we must understand betrayal before we can understand trust; and that we must have a clear understanding of trust before we can understand betrayal.

In this thesis I analyse the concepts of trust, trustworthiness, and betrayal, and explicate the interplay between them. Since the philosophical literature on trust and trustworthiness is more established than that on betrayal, I start by analysing the concepts of trust and trustworthiness. With an understanding of those concepts in place, I develop an account of betrayal that positions me to explain the damage caused to trust by betrayal, and the circumstances under which reasonable trust can be recovered by those who have been betrayed.

In *Chapter 1* I begin by analysing the concept of trust without relying on the concept of betrayal. I demonstrate that a satisfactory philosophical analysis of trust will need to account for a variety of phenomena without compromising the distinction between trust and mere reliance. I then discuss four influential approaches to understanding the concept of trust: i) risk-assessment approaches advocated, for example by Diego Gambetta (1988) and Russell Hardin (1991, 1996, 2002, 2006, 2004); ii) will-based approaches argued for by Annette Baier (1986) and Karen Jones (1996); iii) Strawsonian participant stance approaches, such as those of Richard Holton (1994), Victoria McGeer (2002); Pamela Hieronymi (2008), and Stephen Wright (2009); and iv) trust-responsiveness approaches developed by Jones (1996) and Philip Pettit (1995). While each of these approaches accounts for some features of trust and explains some trust phenomena, I show that they are all unsatisfactory. Each fails to account for some trust phenomena and/or for all of the difference between trust and mere reliance. In light of the limitations of the above approaches, I propose that it is necessary to reconsider the *type* of concept that trust is. Adapting Ludwig Wittgenstein's idea of "family resemblance" (1958, p. 27) and Natalie Stoljar's analysis of the concept "woman" (1995, p. 262), I develop an account of trust as a cluster concept, involving a cluster of different features. On this account, not all instances of trust share common features. Rather, what makes an instance of trust count as trust is that it is sufficiently similar to paradigm cases of trust, which are those that involve features that are more central in trust's cluster. This approach is able to accommodate various different trust phenomena and account for the difference between trust and mere reliance without positing a special relation between trust and betrayal.

Having analysed the concept of trust, in *Chapter 2* my focus shifts to persons whom it is reasonable to trust. I start with the claim that it is at least reasonable to trust the trustworthy. They are, if nothing else, worthy of our trust, and it is prime facie true that we have good reason to trust those who are worthy of it. I discuss accounts of trustworthiness by Carolyn McLeod (2011); Hardin (1991, 1996, 2002); Baier (1986, 1994); Jones (2011); and Pettit (1995), which understand trustworthiness as a matter of competence and commitment to do what another is counting on you to do. I show that it is possible for persons to be relevantly competent and committed without necessarily being trustworthy. To that end I adapt Nancy Potter's (2002) account of trustworthiness as a kind of integrated, virtuous character. Potter explains the kind of character that the trustworthy have by developing a list of requirements for trustworthiness. I argue that these are too stringent and provide moderate versions of her more demanding requirements.

After arguing for a character-based account of trustworthiness, I show that the trustworthy are not the only persons that it can be reasonable to trust. It can also be reasonable to trust persons despite not knowing much about their character. I introduce the concept of "trustability" to distinguish the larger class of those it is reasonable to trust from those who have trustworthy character. On the account I advance here, being trustworthy is just one way of being trustable. Persons can also be trustable because of the character of their wills, because we expect them to be trust-responsive, or because they have various other incentives to uphold our trust. In *Chapter 5* I draw on these various motives for trustability when exploring conditions under which it is possible to recover reasonable trust after betrayal.

I begin *Chapter 3* by showing that trust is not only vulnerable to being damaged by individuals, but can also be damaged by institutions and their role-holders. Given this possibility of damage, concepts of trust, trustability and trustworthiness are relevant to institutional contexts. However, applying them to those contexts is complicated by three challenges. First, while relations among individuals, institutions and role-holders are similar in some respects to interpersonal relations, they may be insufficiently similar to paradigm cases of trust to be instances of the cluster concept of trust themselves. Second, Hardin has argued that seeming instances of trust in institutional contexts should be understood as "quasi-trust" since, he thinks, persons cannot have sufficient knowledge about institutions for reasonable trust (Hardin, 2002, p. 157). Third, at first glance, the character-based view of trustworthiness that I argue for in *Chapter 2* seems inapplicable to institutions. It is unclear whether or not institutions have character. And if they do, it needs to be explained what it means for their character to be integrated and virtuous. I respond to these challenges and argue that it is possible to provide substantive accounts of institutional trust, trustability, and trustworthiness.

When developing my account of institutional trustability, I argue that the condition Onora O'Neill refers to as "intelligent accountability" can encourage institutions and their role-holders to be trustable, while what she calls "active checking" can aid us in knowing whether institutions and role-holders really are competent and committed to doing what we are counting on them to do (O'Neill, 2002b, pp. 58, 77). It turns out that we can have sufficient knowledge about institutions for reasonable trust.

After explaining how institutions can have character, I develop an account of institutional *integrated, virtuous* character that draws on both Justin Oakley and Dean Cocking's (1998) view of virtuous professional roles and on the character-based view of trustworthiness I adopted from Potter (2002) in *Chapter 2*. I argue that trustworthy institutions are characteristically disposed to contribute to human flourishing: the institution's roles, role-holders and policies dispose them to fulfil functions which, when performed well, contribute to human well-being. I use the moderate version of Potter's (2002) requirements that I developed in *Chapter 2* to articulate what it means for institutions to be disposed to perform functions well.

Having provided substantive accounts of trust, trustability and trustworthiness, in *Chapter 4* I develop an account of the concept of betrayal. Drawing on three cases of betrayal, I develop a preliminary analysis of betrayal as a disappointment of normative, "constitutive" expectations that shape particular relational domains. This view of betrayal is similar to Simon Keller's (2007) understanding of disloyalty. Both betrayal and disloyalty involve disappointment of normative expectations. However, I show that it is possible to be disloyal without betraying. Because of that possibility, I argue that betrayal is best understood as a *kind* of disloyalty that involves a failure to uphold normative expectations in a way that undermines special relationships. I conclude *Chapter 4* by reconsidering the special relation between trust and betrayal that is often assumed in analyses of trust. I show that like trusters, those merely relying on others can be vulnerable to betrayal. However, trust, but not mere reliance, can establish the special relationships undermined by betrayal. As such, there is still reason to associate betrayal with trust.

I elaborate on the relation between trust and betrayal in *Chapter 5* by analysing the ways that betrayal can damage trust and explaining how it is possible for victims of betrayal to place trust reasonably again. After presenting three cases of betrayed trust I identify three effects that betrayal can have on trusters: distrust, loss of confidence in knowing who can be trusted, and responses such as anger and hostile emotions. Drawing on these effects I develop an account of the damage that betrayal can do to trust. I show that betrayal can damage conditions for trust; foster conditions for distrust; and strike at its victims' confidence about their judgements of another's trustability, their expectations that others will do what they ought to do, and their background sense of security.

The damage that betrayal inflicts on its victims can make it difficult for them to trust. Unsure of their own judgements, doubtful that others will uphold normative expectations, and aware of their own vulnerability, victims of betrayal may withdraw into self-protective distrust. Given the damage that betrayal can do, I outline the conditions that make it possible for victims of betrayal to reasonably place trust in others once again. I show that forgiveness which responds to a betrayer's expression of remorse is important for recovering reasonable trust in one's betrayer. Using McGeer's view of Strawsonian co-reactive exchange (McGeer, 2011), I provide an account of forgiveness as such a response. When betrayers take responsibility for their actions and express authentic remorse to their victims, the betrayed may have their resentment of the offender mitigated and gradually become more optimistic about trusting her. But betrayal does not only damage a person's trust in their betrayers. The cases I discuss at the beginning of *Chapter 5* show that betrayal can cause what I term "collateral distrust" in others more broadly. I conclude *Chapter 5* by suggesting ways in which individuals, institutions and institutional role-holders can scaffold a victim of betrayal's self-confidence as a truster and so help a victim who experiences collateral distrust to recover reasonable trust. I show that this scaffolding work can be done by the "moral community", insofar as they affirm the judgements of trustability and the normative expectations that the victim had before being betrayed.

My analysis in this thesis contributes to our understanding of the landscape of moral agency. Hopefully, my contribution will encourage further debate, and greater understanding, of betrayal, its relation to trust, and its impact on human flourishing.

Chapter 1: Trust as a Cluster Concept

Section 1: Introduction

Whatever trust is, it is integral to many parts of life. We trust partners not to be cheaters and friends to keep our secrets. We may trust acquaintances with personal information such as phone numbers and addresses, and trust strangers not to mislead us when we ask for directions. We may even trust enemy soldiers not to shoot when our white flag of surrender is raised. While its phenomena are present in a variety of domains and situations, trust remains, putatively, distinct from related concepts such as mere reliance.[1] This is because it seems possible to rely without trusting. For example, consider the plight of the working class in Victorian England.[2] In *The Condition of the Working Class in England in 1844*, Friedrich Engels explains that the working class had little access to good food (Engels, 2008, p. 68). Often receiving their wages on Saturday evening, workers were unable to deal with market vendors until after middle class shoppers had purchased the better products. Even if some of those goods remained available, it is likely that workers would not have had the funds necessary to purchase them. Add to this the fact that many vendors would sell adulterated products to workers who were either unaware that the goods had been tampered with or could not afford anything else. Ground rice was added to sugar; finely sifted dirt was mixed through cocoa; and chalk was used to stretch flour (Engels, 2008, pp. 69-70).

> ...The poor, the working-people, to whom a couple of farthings are important, who must buy many things with little money, who cannot afford to inquire too closely into

[1] It is commonly accepted amongst philosophical analyses that trust is distinct from mere reliance. This claim is made most directly by Annette Baier and Carolyn McLeod. Baier writes, “We can still rely where we no longer trust” (Baier, 1986, p. 234). And McLeod says, “Although people who monitor and constrain other people's behavior and do not allow them to prove their own trustworthiness may rely on others, they do not trust them” (McLeod, 2011).

[2] My thanks to Cynthia Townley for this example.

> the quality of their purchases, and cannot do so in any case because they have had no opportunity of cultivating their taste—to their share fall all the adulterated, poisoned provisions. (Engels, 2008, p. 71)

As the above quote from Engels points out, some workers would have consumed adulterated products unknowingly. But others would have purchased them despite suspicion of the vendor and her products. They would have had to rely on the vendors, despite not trusting them. Whatever trust is then, it is an integral and broad, *but distinct*, part of life.

In this chapter I show that a satisfactory philosophical analysis of the concept of trust will need to account for a variety of phenomena without compromising the distinction between trust and mere reliance. Having shown that, I argue that trust is best understood as a cluster concept. As such, the concept of trust involves a broad group of features, types and characteristics of phenomena while remaining distinct from mere reliance. The result is an understanding of trust consistent with the kind of world we live in; one where persons count on others in different ways and can be disappointed, let-down and even betrayed.

In section 2 I present four examples that I take to be characterised by trust in uncontroversial ways. I use these examples to illustrate the variety in trust phenomena and to motivate my analysis. In section 3 I identify and explain variations in trust phenomena based on those examples. I show that trust occurs in a range of different circumstances including relations with intimates, acquaintances and strangers, and relations that are iterated or merely one-off interactions. I explain that vulnerability is a central feature of these relations but that trusters can be vulnerable in various ways. They can be practically vulnerable to having their entrusted goods and/or themselves harmed by another. But they can also be vulnerable to what I will call “moral harms” like ill will, deception and betrayal. In the examples I introduce, acceptance of vulnerability is a related central feature of trust but one which also can vary. Persons can accept vulnerability to others confidently or they can do so despite a lack of confidence. In addition to variations in trust phenomena evident in my examples, in section 3 I also accept three characteristics of trust identified by Jones: trust is contrary but not contradictory to distrust; trust cannot be willed; and trust gives rise to beliefs that are abnormally resistant to evidence (K. Jones, 1996, p. 15). A satisfactory understanding of trust will need to account for various features, types and characteristics of trust *and* for the fact that it is distinct from mere reliance.

In section 4 I consider four influential approaches taken toward the concept of trust and argue that each is unsatisfactory, either because it is unable to account for various trust phenomena or because it does not account for at least some difference between trust and mere reliance. I conclude section 4 by proposing a reconsideration of the type of concept trust is often taken to be.

In section 5 I use Ludwig Wittgenstein's idea of family resemblances (Wittgenstein, 1958, p. 27) to identify two types of concepts: those amenable to analysis in terms of necessary and sufficient conditions and those identifying similarities between things but which do not have features common to all their instantiations. I follow Natalie Stoljar in understanding this latter type of concept as a "cluster" and argue that trust is best understood as such (Stoljar, 1995, p. 283). In addition to being able to account for the various different phenomena that can be characterised as trust, and the fact that trust is distinct from related concepts, taking a cluster approach aids explanation of border-line cases and makes sense of common usage of trust language.

Section 2: Four Examples of Trust

To illustrate the variety of domains where trust occurs, I begin by introducing four examples of trust. I have chosen examples from *interpersonal* domains. This is because I think they are clearly characterised by trust. While it may seem appropriate to talk about other parts of human experience, such as relying on natural phenomena, in terms of trust, it is less clear to me that such language should be taken at face value.[3] Rather, I think statements such as, "I trust the sun to rise tomorrow" should be understood as metaphorical. At the outset it is difficult to explain why this is the case. But after arguing for my view of trust as a cluster concept I will be able to explain why such talk should not be taken at face value. For the time being I will motivate my analysis with interpersonal examples that, I think, are less questionably characterised by trust.

Each example I have chosen involves a different degree of intimacy between the person(s) doing the trusting – i.e. the truster(s) – and the person(s) being trusted – i.e. the trustee(s). They range from relationships between closely connected individuals through strangers to enemies, and have been ordered starting with more intimate relations.

The Co-Parent Relationship[4]

The relationship between co-parents is one characterised by a high level of intimacy and trust. In most cases co-parents live closely with each other, have extensive shared knowledge and experiences and count on each other for the raising of their children. But a level of trust can persist even after that intimacy changes. Co-parents that are separated can still have a relationship involving trust. And this can be true even if that relationship exists in a context of conflict. When co-parents are separated, their child's care will usually be provided by one parent at a time in his or her own home. This results in parents sharing responsibility for the care of their child but often giving that care individually. Parents in such circumstances may exercise a level of trust when leaving their child in the discretionary care of their co-parent. If separated co-parents feel animosity toward each other, any trust involved in leaving their child in the other's care will be different than if their relationship is

[3] Philip Pettit has claimed that it is appropriate to refer to reliance on natural phenomena as trust (Pettit, 1995, p. 203). I will return to explain his view in section 1.3.4.

[4] I take this example from Baier's mention of trust between separated co-parents. "When one trusts one's child to one's separated spouse, it is all aspects of the child's good as a developing person which are entrusted to the other parent's care" (Baier, 1986, p. 238).

more amicable. But regardless of the feelings that separated co-parents have toward each other, sharing responsibility for the care of a child often involves some amount of trust.

I have chosen to focus on the trust between separated co-parents rather than parents who are unseparated in order to isolate trust from situations over-determined by affectionate ties. Of course, some separated co-parents will continue to hold affection for each other; but in many cases affection does not persist. Rather, such parents may entrust each other with the care of their child simply because they have been ordered to do so by a court of law.[5] Or they may do so because, despite their hard feelings, they take the other to be someone who can be trusted. Regardless, the relationship between co-parents can continue to be one involving trust even after separation.

The "Trust Fall" Exercise

In "Deciding to Trust, Coming to Believe" Richard Holton explains a common team building activity often referred to as a "trust fall" exercise (Holton, 1994, p. 63). In the activity, one person is blind-folded and made to stand in the middle of a circle formed by other teammates. The blind-folded person is turned around so as to make her disoriented and unable to judge how far she is from those standing around her. Then the teammates encourage the blind-folded person to fall backwards into their arms.

In the "trust fall" exercise the blind-folded person must trust that her teammates will not, literally, let her down. That is, she must trust that the teammates will catch her when she falls. As Holton explains,

[5] If separated co-parents entrust the care of their child to each other because of a judicial directive rather than agreeing to an arrangement on their own, it is more likely that any trust between them will involve resentment, ill will and suspicion. The presence of these attitudes may call into question the involvement of trust in their relationship. And indeed, not all separated co-parents will trust each other. But I think such parents can still be in a position where they trust each other with the care of their child. And I think that remains the case whether they do so of their own choice or because they have been ordered to do so by a court. If they entrust their child to each other's care only because of a judicial directive, I think it is unlikely that their relationship will be characterised by trust. But I think it is possible for trust to be present between parents even when a court order is a reason they entrust the care of their child to each other. They may be ordered to entrust their child but also consider each other to be persons who can be trusted to some degree.

> ...with your arms by your sides and your legs straight, you let yourself fall. You let yourself fall because the others will catch you. Or at least that is what they told you they would do. You do not know that they will. You let yourself fall because you trust them to catch you. (Holton, 1994, p. 63)

The Peruvian Artist

In "Trusting People", Baier tells a story of trust between strangers (Baier, 1992, p. 139). A Peruvian artist selling paintings at a fair in the US is approached by a couple who offer to purchase a piece of his work on the condition that he personally delivers it to their home. The Peruvian is working his art stall alone and so is faced with a decision to either close the stall and deliver the art work or keep the stall open but miss the sale. He chooses neither of these options. Instead he solves the dilemma by proposing that one of the patrons watch his stall while the other guides him to their home (Baier, 1994, p. 186). I take the artist's decision and actions to express trust in the patrons. He leaves them in a position where they could harm him and his stall.

The Christmas Truce

We do not only trust intimates, colleagues or teammates and strangers. At times we also trust enemies or those with whom we are otherwise in conflict with. Baier comments on this possibility by simply saying, "Trust, the phenomenon we are so familiar with that we scarcely notice its presence and its variety, is shown by us and responded to by us, not only with intimates but with strangers, and even with declared enemies" (Baier, 1986, pp. 233-234). Baier does not provide an example of trust being placed in enemies, but what has been referred to as "The Christmas Truce of 1914" is such a case (Karan, 2010; Rees, 2009; Richards, 2006).

On December 24th 1914, along sections of the western front from St. Yves to Neuve Chappelle, an informal truce was agreed upon between German and British soldiers. Facts and myths about The Christmas Truce have blended together, but it appears that German soldiers initiated the truce by sending some chocolate and a note, proposing a cease-fire during Christmas day. British soldiers responded by sending back a gift of tobacco and a note agreeing to the cease-fire. Frank Richards has written about his experience in this case of trust between enemies.

> On Christmas morning we stuck up a board with 'A Merry Christmas' on it. The enemy had stuck up a similar one. Platoons would sometimes go out for twenty-four hours' rest - it was a day at least out of the trench and relieved the monotony a bit - and my platoon had gone out in this way the night before, but a few of us stayed behind to see what would happen. Two of our men then threw their equipment off and jumped on the parapet with their hands above their heads. Two of the Germans done the same and commenced to walk up the river bank, our two men going to meet them. They met and shook hands and then we all got out of the trench.
>
> Buffalo Bill [the Company Commander] rushed into the trench and endeavoured to prevent it, but he was too late: the whole of the Company were now out, and so were the Germans. He had to accept the situation, so soon he and the other company officers climbed out too. We and the Germans met in the middle of no-man's-land. Their officers was also now out. Our officers exchanged greetings with them. One of the German officers said that he wished he had a camera to take a snapshot, but they were not allowed to carry cameras. Neither were our officers.
>
> We mucked in all day with one another. They were Saxons and some of them could speak English. By the look of them their trenches were in as bad a state as our own. One of their men, speaking in English, mentioned that he had worked in Brighton for some years and that he was fed up to the neck with this damned war and would be glad when it was all over. We told him that he wasn't the only one that was fed up with it. We did not allow them in our trench and they did not allow us in theirs.
>
> The German Company-Commander asked Buffalo Bill if he would accept a couple of barrels of beer and assured him that they would not make his men drunk. They had plenty of it in the brewery. He accepted the offer with thanks and a couple of their men rolled the barrels over and we took them into our trench. The German officer sent one of his men back to the trench, who appeared shortly after carrying a tray with bottles and glasses on it. Officers of both sides clinked glasses and drunk one another's health. Buffalo Bill had presented them with a plum pudding just before. The officers came to an understanding that the unofficial truce would end at midnight. At dusk we went back to our respective trenches. [...]
>
> Just before midnight we all made it up not to commence firing before they did. At night there was always plenty of firing by both sides if there were no working parties or patrols out. Mr Richardson, a young officer who had just joined the Battalion and was now a platoon officer in my company wrote a poem during the night about the Briton and the Bosche meeting in no-man's-land on Christmas Day, which he read out to us. A few days later it was published in *The Times* or *Morning Post,* I believe.
>
> During the whole of Boxing Day [the day after Christmas] we never fired a shot, and they the same, each side seemed to be waiting for the other to set the ball a-rolling. One of their men shouted across in English and inquired how we had enjoyed the beer. We shouted back and told him it was very weak but that we were very grateful for it. We were conversing off and on during the whole of the day.

> We were relieved that evening at dusk by a battalion of another brigade. We were mighty surprised as we had heard no whisper of any relief during the day. We told the men who relieved us how we had spent the last couple of days with the enemy, and they told us that by what they had been told the whole of the British troops in the line, with one or two exceptions, had mucked in with the enemy. They had only been out of action themselves forty-eight hours after being twenty-eight days in the front-line trenches. They also told us that the French people had heard how we had spent Christmas Day and were saying all manner of nasty things about the British Army. (Richards, 2006)

Soldiers involved in The Christmas Truce needed to trust their leading officers' acceptance of the truce with the enemy officers. And they had to trust the opposing enemy officers and soldiers to uphold the truce and not harm them upon leaving their trenches. I take this example to be an instance of trust between enemies that are strangers. But it is also possible to have trust in the context of conflict between intimates, or persons like teammates or co-workers, where the degree of intimacy between those involved is somewhere between that of strangers and intimates. For example, as I have already pointed out, the separated co-parents may have deep animosity towards each other, and so be enemies in a sense, while still trusting each other's care of their shared child. Teammates or co-workers may also trust each other despite hard feelings. For example, despite conflict co-workers may be able to trust each other to do what needs to be done at work in order to be successful as a company. Likewise, sports teammates may harbour hard feelings toward each other but still trust each other to try and defeat an opposing team.

Each of the above four examples is characterised by trust, but each involves a significantly different type of interaction and relationship. Attention to both the similarities and the differences between these cases will enable identification of features and types of the concept of trust.

Section 3: Features, Types, and Characteristics of Trust Evident in the Phenomena

In this section I identify and explain similarities and differences in the above four examples and accept three characteristics of trust identified by Karen Jones. Those characteristics are that: trust is contrary but not contradictory to distrust[6]; trust cannot be willed; and trust can lead to beliefs that are, to use Jones' phrase, "abnormally resistant to evidence" (K. Jones, 1996, p. 15). My analysis proposes a range of features, types and characteristics for which a satisfactory understanding of the concept of trust should account.

To begin with, all the above cases involve some type of *counting on another, or others*, for something. For example, the separated co-parents count on each other to care for their child. The "trust fall" participants count on their teammates when it is their turn to fall. The Peruvian artist counts on one patron to watch his stall while he is away and he counts on another patron to guide him to the patrons' home. And the Christmas Truce soldiers count on their opposing soldiers to uphold the truce and not shoot them when they leave their trenches. Counting on another or others seems to be a central feature of what it means to trust. But it is not distinctive of trust since it is possible to count on someone and be merely relying on them without trusting them. Still, satisfactory understandings of trust will need to account for it.

The examples also involve varying *degrees of intimacy* between the parties to the trust relationship. The *un*separated co-parent relationship involves trust between intimates. But as I showed in section 2, the relationship between separated co-parents can also involve a level of intimacy. In contrast, the Peruvian artist and the enemy soldiers in the Christmas Truce example trust persons that are mostly strangers to them. The artist has only just met those he trusts and the soldiers trust people they have never seen. The "trust fall" example is a case of trust in those who are somewhere between strangers and intimates. I will call this trust in

[6] While I follow Jones on this point, it was first made by Govier. She writes, "Trust and distrust are contraries, not contradictories: we may neither trust nor distrust" (Govier, 1992b, p. 18). Jones also cites Govier on this point (K. Jones, 1996, p. 15). Edna Ullmann-Maralit has made a similar claim in saying, "Trust and distrust, while mutually exclusive, are not mutually exhaustive" (Ullmann-Margalit, 2004, p. 60).

"acquaintances" and will include in it a large class of trust occurring between teammates, co-workers and friends that are not closely connected. Satisfactory explanations of trust will be able to account for the variation in degree of intimacy that can occur in trust.

The *duration of interaction* also varies in the examples I have provided. Some of the interactions are "one-off" while others are iterated. For example, the Peruvian artist will probably not interact with his patrons after delivering the artwork and completing his transaction with them unless the patrons become return customers. In contrast, the "trust fall" participants are probably involved in the exercise because they will be working together in the future and as a result, want to grow the sense of unity and trust between them. And regardless of the developments in their own relationship, the interaction of the separated co-parents will be iterated as long as they, and their child, are alive. The extent of their parenting, and so the extent of their interaction with each other, will change as their child becomes more independent, but they will probably need to interact to some extent as long as their child requires parenting.

The Christmas Truce example is an interesting case with regard to duration of interaction. Unlike the Peruvian artist's "one off" trust in the patrons, the warring soldiers knew that they would continue to interact with each other for some time after the truce had finished. While they may have continued to relate to each other in ways characterised by trust, they would interact, albeit in a context of conflict, until the war was over. Based on the various durations of interaction in these examples, satisfactory understandings of trust must account for involvement that is "one-off" and for that which is expected to continue, whether in a context of peace or conflict.

Each of the above cases involves *vulnerability to being let-down and/or having one's valued goods harmed.* In the separated co-parent relationship each parent can be vulnerable to having their child harmed by the other. The participants in the "trust fall" example are vulnerable to not being caught by their teammates when they fall and experiencing physical injury as a result. The Peruvian artist is vulnerable to having art work stolen from his stall while he is away. And a severe vulnerability to physical harm is evident in the Christmas Truce example: the warring soldiers are vulnerable to their opponents breaking the truce and firing their weapons at them. In leaving their trenches the soldiers risk their very lives. I will refer to trusters that are in a position where they may be let-down and/or have their valued goods damaged as being *practically vulnerable*.

In addition to practical vulnerability, these examples involve persons who are, what I will call, *morally vulnerable*. They are vulnerable to deception, betrayal and other acts expressing attitudes of ill will or indifference. I am referring to such vulnerability as "moral" because it has to do with the attitudes members of the moral community have toward one another. Another's attitudes of goodwill, ill will or indifference matter to us and can affect our response toward their behaviour. I take this idea from P.F. Strawson.

> If someone treads on my hand accidentally, while trying to help me, the pain may be no less acute than if he treads on it in contemptuous disregard of my existence or with a malevolent wish to injure me. But I shall generally feel in the second case a kind and degree of resentment that I shall not feel in the first. If someone's actions help me to some benefit I desire, then I am benefited in any case; but if he intended them so to benefit me because of his general goodwill towards me, I shall reasonably feel a gratitude which I should not feel at all if the benefit was an incidental consequence, unintended or even regretted by him, of some plan of action with a different aim. (Strawson, 1974, p. 5)

In addition to risking the harm of their valued goods, the trusters in the cases I have introduced take the risk of being responded to out of ill will or indifference. In the separated co-parent case, one parent's caring for the shared child in a way that the other disapproves of may simply be a mistake or difference in parenting style. For example, one co-parent may allow their child to play a sport that the other co-parent considers too dangerous. But such parenting could also express indifference, ill will or spite. Moral vulnerability is also evident in the Peruvian Artist case. Since that case involves trust between strangers, a failure of the patron to care well for the artist's stall would probably not be taken to express spite toward the artist. But the patron's poor care could be taken to express indifference toward the artist and his goods. And as already noted, participants in the "trust fall" exercise are vulnerable to bodily harm should their teammates not catch them. If the teammates try to catch the faller but fail in their attempt, their failure will probably just be taken to express a physical weakness on their part. But the blind-folded participant is also morally vulnerable. Should the teammates tell the blind-folded person to fall and then not even try to catch her; their actions can be taken to express a negative attitude toward the faller (e.g. dislike, contempt.) Satisfactory understandings of trust will need to account for the practical *and* moral vulnerability involved in trust.

One more point needs to be made about moral vulnerability. Vulnerability to betrayal may be one type of moral vulnerability. For example, a "trust fall" participant who is not caught by her teammates may take the teammates' actions to be intentional and feel betrayed. But the wounded "trust fall" participant may take her teammate's actions to be intentional and feel anger or have some other negative reaction *but not feel betrayed*. This suggests that vulnerability to betrayal is not the whole of moral vulnerability. This is important to point out since it is common to distinguish trust from mere reliance on the basis that trust involves vulnerability to betrayal while mere reliance does not. But that is not my claim here. The moral vulnerability I am identifying is not just vulnerability to betrayal. Rather, it is vulnerability to ill will or indifference. Betrayal can express ill will or indifference, but it is not the only thing that can do so.

The four examples in section 2 involve *acceptance of vulnerability* on the part of the truster insofar as none of those involved resist being vulnerable to their respective trustees. The separated co-parents leave their child in each other's care, the person in the centre of the "trust fall" circle begins to fall, the Peruvian artist leaves his stall in the patron's care and the

Christmas Truce soldiers leave their trenches.

In most of the examples in section 2, acceptance of vulnerability is explicit and occurs voluntarily at a point of decision. For example, faced with a decision to fall or not, the blind-folded “trust fall” participant chooses to accept the risk of falling. The Peruvian artist also accepts vulnerability to his patrons voluntarily and at the point of a business decision. Wanting to make a sale, he takes the risk involved and proposes that one patron guard his stall while the other leads him to their home. Voluntary acceptance of vulnerability is also evident in the Christmas Truce example. At some point the soldiers involved in the truce would have made a decision to accept their vulnerability to the enemy and leave their trenches.

But not all vulnerability is accepted so explicitly or because of a voluntary decision. This is evident in some types of co-parent relations. Some separated co-parents harbouring hard feelings may only accept any vulnerability to each other after a decision to do so – they might harbour some suspicions about their ex-partner but after some deliberation come to accept vulnerability and leave the child with the co-parent. But, acceptance of vulnerability between co-parents can be implied if they have come to trust each other over time and that trust has not been called into question. It is likely that, from before they were co-parents together, the trust between them grew incrementally as they came to interact more together and know each other better. Rather than explicitly choosing to accept vulnerability at any specific points in time, the acceptance of vulnerability by such unseparated co-parents is *implied* as they continue on in their relationship together. They do not identify the risks involved in trusting each other and then choose to accept vulnerability anyway. They unreflectively trust each other, at least until that trust is called into question. As Baier has pointed out, “We inhabit a climate of trust as we inhabit an atmosphere and notice it as we notice air, only when it becomes scarce or polluted” (Baier, 1986, p. 234). As the examples in section 2 show, not all trust is so unreflective. But when it is, acceptance of vulnerability will most likely be implied rather than explicit and voluntary. Satisfactory understandings of trust will need to allow for acceptance of vulnerability that is explicit and voluntary and for that which is implied.

In cases where vulnerability is accepted voluntarily, it is plausible that the confidence with which one accepts that vulnerability will vary. For example, the Peruvian artist might have not thought twice about his decision to leave his stall in the care of one patron and go with the other to deliver his artwork. Or he might have trusted them despite being anxious and harbouring doubts about whether the patrons would take advantage of him and his stall. He may have just chosen to trust despite his doubts because he needed to make the sale. It is even possible for persons to trust despite having good reason to think they will be let-down by another. For example, consider a parenting relationship different to that between co-parents: the relationship between a parent and her teenage son who has just received a driving license. If the teenager requests permission to use the parent's car on a Saturday night, the parent might refuse because of great doubts about the teenager's current responsibility when it comes to cars. But she may also decide to agree to the use of the car despite those doubts in the hope that treating the teenager as if he were responsible might encourage him to become a more responsible pre-adult. That is, the parent may trust, not because the teenager is currently worthy of that trust but to encourage him to be worthy of it.[7]

Trudy Govier, H.J.N Horsburgh and Carolyn McLeod call the type of trust that the parent in the above example places in the teenage driver "therapeutic trust" (Govier, 1992b; Horsburgh, 1960; McLeod, 2011).[8] Just as physical therapy is meant to train a part of the body to perform some motion, the parent's trust is meant to train the teenager to be responsible and responsive to the fact that she is counting on him. I do not think "therapeutic

[7] This example of the parent entrusting the son or daughter with the family car is also used by Victoria McGeer (McGeer, 2008, p. 241); and by Karen Jones (K. Jones, 2004, p. 5).

[8] Horsburgh first coined the term "therapeutic trust" in "The Ethics of Trust" (Horsburgh, 1960, p. 346).

trust” is the best term for the phenomenon being identified. In addition to connotations of *training* something or someone, the term “therapeutic” has connotations of *repairing* something that has been broken. For example, persons may receive physical therapy to repair bodily injuries or seek help from a therapist to heal emotional or psychological trauma. But interaction referred to as therapeutic trust does not always involve restoring trust that has previously been damaged. It may instead involve building trust that had previously not been present. This is the case in the relationship between the parent and the teenage son. The parent’s act of entrusting the son with the family car is not meant to repair damaged trust but rather to grow, or develop, a new type of trust between the parent and the son. It scaffolds the trustee’s agency and aims to encourage him to behave in ways that make him worthy of trust. And it can develop a relationship into one of trust whether that means repairing broken trust or establishing trust in the first place. For this reason I think it is better to refer to interaction aimed at repairing or establishing a trust relationship as “developmental” rather than “therapeutic” trust. Satisfactory understandings of trust will need to allow for trustee acceptance of vulnerability that is confident and for less confident acceptance as occurs in developmental trust.

A satisfactory understanding of trust should also be able to account for three characteristics Jones has identified but which are not evident in the above examples. Trust is contrary but not contradictory to distrust; it cannot be willed; and trust gives rise to beliefs that are, “abnormally resistant to evidence” (K. Jones, 1996, p. 15). I will explain these three characteristics in section 4.2 when presenting Jones’ argument against understanding trust on an entrusting model.

In summary, the examples in section 2 show that trust can be placed in strangers, acquaintances or intimates; trust can be shown in one-off interactions and in iterated relationships; trust can involve practical and moral vulnerability; and trust can involve confident acceptance of vulnerability or acceptance of vulnerability despite the presence of relevant doubts. Satisfactory understandings of trust should be able to account for these features and types of trust as well as the characteristics identified by Jones without compromising the distinction between trust and mere reliance.

Section 4: Influential Approaches to the Concept of Trust and Their Limitations

The list of features, types and characteristics I identified in section 3 is extensive. It will not be easy for an approach to trust to account for all of the variety in phenomena I have identified without compromising trust's distinctiveness. And yet that is what needs to be done if we are to have an understanding of the concept of trust that matches what goes on in the world.

In this section I discuss four influential approaches to trust: trust as risk-assessment; trust as reliance on the goodwill of another toward you; trust as a participant stance toward another; and trust as an expectation of trust-responsiveness. I argue that each approach is unsatisfactory. They are either unable to account for some trust phenomena, or do not sufficiently explain the difference between trust and mere reliance. For an approach to be insufficient as an understanding of trust it need only fail to account for one type of phenomena, or for the distinction between trust and mere reliance. Therefore, I do not need to identify each and every way in which the approaches I consider fail to be satisfactory. Instead, I highlight limitations that, I think, show the insufficiency of each approach most clearly. I begin by considering risk-assessment approaches because they respond to one of the most common elements occurring in trust phenomena – the risk involved in counting on another.

Section 4.1: Risk-Assessment Approaches to Trust

Trust is risky. When we trust others we may be let down, betrayed or have ourselves, our goods and/or our trust damaged in some other way. This is evident in the phenomena I identified in section 2. Co-parents risk having their child harmed by each other. The Peruvian Artist risks his artwork being stolen while he makes his delivery. The blind-folded participant in the "trust fall" exercise risks falling to the ground. And the enemy soldiers in The Christmas Truce risk their very lives. And yet, risks are taken in each of these examples.

Some philosophers have made risk-taking behaviour central to their understanding of the concept of trust. They take trust to be interacting with another because one thinks the risk of doing so is acceptable. I will follow Jones and McLeod in referring to this understanding as a "risk-assessment" approach to trust (K. Jones, 1999; McLeod, 2011).

Diego Gambetta's understanding of trust is a good example of a risk-assessment approach.[9] On Gambetta's view, trusters cannot be certain how others will behave. But despite that lack of certainty trusters think it probable that others will behave in a given way. Gambetta writes,

> When we say we trust someone or that someone is trustworthy, we implicitly mean that the probability that he will perform an action that is beneficial or at least not detrimental to us is high enough for us to consider engaging in some form of cooperation with him. (Gambetta, 1988, p. 217)

Gambetta's risk-assessment approach can be made to fit most trust phenomena discussed in sections 2 and 3. First, it is possible to tell a story about assessing the risk of counting on intimates, acquaintances and strangers. Intimate co-parents might be said to be cooperating because they think it is probable that the other will care well for their child. The blind-folded person in the "trust fall" exercise may just fall because she thinks the risk of doing so is sufficiently low. She may have some reason to think the others will catch her. Or she may take the risk and fall because she is not very concerned with the consequences of falling to

[9] Jones first identified Gambetta's view as a risk-assessment account (K. Jones, 1999, p. 68).

the ground. She may just think she will not be badly hurt if the teammates fail to catch her. The Peruvian artist could be said to be counting on his patrons because he thinks the risk of them harming him or his art stall is sufficiently low. And part of the reason the soldiers in The Christmas Truce left their trenches may have been that they thought the risk of the enemy breaking the truce had been decreased to an acceptable level.

Second, the Peruvian artist and the separated co-parenting examples show that it is possible to assess risk in one-off interactions as well as in iterated relationships. The Peruvian can assess the risk of counting on the patrons though he may never interact with them in the future. And the separated co-parents can assess the risk of entrusting their child to each other's care knowing that they will continue to interact. While a risk-assessment story can be told about both one-off interactions and iterated relationships, some risk-assessment views, such as Russell Hardin's (1991, 1996, 2002, 2006, 2004) encapsulated interest account, fit iterated relationships better than one-off interactions. This will be made clearer when I explain Hardin's view below. My current claim is that risk-assessment approaches in general can account for trust occurring in one-off interactions and in iterated relationships.

Third, it is possible to assess the risk of harm whether practical or moral. For example, consider again the Peruvian artist. In assessing whether to leave his stall in the patron's care, the artist can assess what he thinks the chances are that the patron will harm him practically by mistreating his stall. But at first glance, it is unclear whether it is appropriate to talk about the artist assessing the risk of *moral* harm in counting on the patrons. It seems strange to think that the artist would assess whether practical harm done by the patrons would be an expression of ill will or indifference toward him. But I think this seeming strangeness can be accounted for by the fact that when we trust others whose actions can be taken to express their states of will toward us, the risk of moral harm is always present. So, it is at least possible for the artist to assess the risk of moral harm in trusting the patrons. He might assess that the risk of them having ill will or indifference toward him exists but accept that risk anyway. He may just think the patrons seem like friendly people and so are unlikely to have ill will toward him.

Lastly, risk-assessment approaches can account for trust that involves confident acceptance of vulnerability. On such approaches trusters remain vulnerable to harm but take the risk of that harm to be acceptable. And so, trusters can accept vulnerability to harm confidently precisely because they are unopposed to the risk of doing so. For example, separated co-parents may leave their child in the other's care because they think the risk involved in doing so is low. If the parents thought about it, they would know that there is always the risk that the other will harm their child through negligence. And in risking that harm they also risk being hurt themselves as one who cares for the child. They are practically vulnerable to having their beloved child harmed and they are morally vulnerable to the damage the child's harm would mean for themselves. They remain vulnerable to each other in those ways but accept that vulnerability because they take the risk of doing so to be low.

While able to account for the acceptance of practical and moral vulnerability, risk-assessment approaches are unable to explain why trusters would be morally vulnerable in the first place. It is unclear why merely interacting with someone because you take the risk of doing so to be acceptable should make one vulnerable to deception, betrayal and other expressions of ill will or indifference. This limitation identifies a need to look beyond risk-assessment when explaining all of the concept of trust.

The most significant challenge facing risk-assessment approaches is that they tend to conflate trust and mere reliance. This is because it is possible to rely on another because the risk of doing so is acceptable without trusting them. Recall the example I presented in the introduction to this chapter of Victorian workers purchasing adulterated goods. Despite thinking a vendor has tampered with her products, a worker may take the risk of purchasing adulterated goods to be acceptable given his situation. He may just think that adulterated food is better than no food at all. And yet, he does not trust the vendor. He takes the risk of buying from her to be acceptable despite his suspicions about her and her products.[10]

When she first referred to some views of trust as risk-assessment approaches, Jones pointed to one reason why they conflate trust and mere reliance: because risk-assessment views tend not to specify reasons for thinking another will behave in a given way. "One might think an agent will perform that action out of fear or stupidity, because it coincides with her own self-interest, or because she has, and wishes to display, goodwill toward those who are counting on her" (K. Jones, 1999, p. 68). Hardin develops a risk-assessment approach that *does* specify reasons for thinking another will behave in a given way (Hardin, 1991, 1996, 2002, 2006, 2004). But, as I will show, Hardin's view of trust is also unable to account for the difference between trust and mere reliance.

In its most basic sense, Hardin's view is that trust involves counting on another because one has reason to think the other will do what one is counting on them to do (Hardin, 2002). Most of Hardin's account is taken up with analysing the best reasons persons have for expecting others to do what they are counting on them to do. Hardin takes personal interests to be the most consistent motivator of agents and so argues that a person has the most reason to count on others when he thinks those others have encapsulated his interests into their own (Hardin, 1991, p. 189; 2002, pp. 1, 29).

The encapsulation of interests involved in Hardin's view is more than a matter of having interests that coincide. Rather, when a person encapsulates another's interests, he

[10] It may be thought that the Victorian worker example is not a case of interaction based on risk-assessment because the worker has little recourse but to buy from the vendor. It might be thought that the worker does not buy from the vendor because he thinks the risk of doing so is acceptable, but because he has no other choice. But I think this critique conflates *acceptable* risk with *low* risk. Risks that are acceptable are not always low. For example, the risk of a soldier dying in battle is high, but to many patriotic citizens that is an acceptable risk. For the Victorian worker, the risk of buying adulterated goods was also high but, given his situation, acceptable. So I take the Victorian Worker example to be an instance of someone relying on another because the risk of doing so is acceptable but not trusting the other.

makes them his own. For example, it is in the interest of both separated co-parents that their child be well cared for. Their interests with regard to the child's well-being coincide. But that does not mean that each parent has made the other's interests their own. They each care about doing what is best for the child rather than for the other parent. And that is especially true if the co-parents are in conflict with each other. On Hardin's view, for the co-parents to encapsulate each other's interests they would need to make each other's interests their own *because of something to do with each other* (Hardin, 2002, pp. 3, 11). Specifically, Hardin holds that people encapsulate the interests of others because they want their relationships with those others to continue.

> ...I trust you because I think it is in your interest to attend to my interests in the relevant matter. This is not merely to say that you and I have the same interests. Rather, it is to say that you have an interest in attending to my interests because, typically, you want our relationship to continue. (Hardin, 2002, p. 4)[11]

Hardin's encapsulated interest view is essentially a risk-assessment approach specifying reasons for accepting the risk of interacting with another.[12] Persons have reason to accept risk when they believe another has encapsulated their interests. And they have reason to believe that encapsulation has occurred, or will occur, if the other is motivated to continue in relationship with them.

Despite specifying reasons for accepting the risk of interacting with another, Hardin's view is still unable to account for the difference between trust and mere reliance. It is possible to rely on another because she has encapsulated your interests into her own while not trusting her. This can be seen in an example provided by Nancy Potter and also discussed by McLeod. Potter describes a situation in which a male boss treats a female employee "well" because he thinks legal sanctions will be brought against him if he does otherwise (Potter,

[11] Hardin puts the same point elsewhere this way, "...I trust you because I think it is in your interests to take my interests in the relevant matter seriously in the following sense: You value the continuation of our relationship, and you therefore have your own interests in taking my interests into account" (Hardin, 2002, p. 1).

[12] McLeod also presents Hardin's encapsulated interest account as a risk-assessment approach to trust (McLeod, 2011) .

2002, p. 5). McLeod extends Potter's example by asking us to imagine that the boss is not only motivated by fear of legal sanctions but also by a desire to keep the female employee around so that he can daydream about having sex with her (McLeod, 2011).

It is possible to tell a story about the above example in terms of encapsulated interests. Because of his fear of legal sanctions, the boss' interests in treating the female employee well may correspond with the employee's interests. For example it is in the boss' interests not to fire the employee and it is in her interests, at least to some extent, to keep her job. But as I explained above, encapsulated interests are not just interests that correspond. This is why McLeod's extension of Potter's example is important. In wanting to keep the female employee around so he can daydream about having sex with her, the boss has reason to sustain his relationship with the employee. And so he has reason to make the employee's interests his own—to encapsulate her interests into his own.

On Hardin's view, if the female employee counts on the boss for job security because she believes the boss has encapsulated her interests into his own, she qualifies as trusting the boss. But I think that if the employee knew the boss encapsulated her interests for such abusive reasons, she might continue to rely on him for job security but she would not trust him. This outcome suggests that even when risk-assessment approaches specify reasons for thinking another will behave in a given way they are unsatisfactory.

The three approaches to trust that I discuss below are promising insofar as they are not based on risk-assessment: trust as reliance on the goodwill of another; trust as an expectation of trust-responsiveness; and trust as a participant stance. Before explaining these approaches I want briefly to mention a counter-argument to my claim that they are not founded on risk-assessment. It could be argued that trust is interacting with another because one has assessed that the risk of doing so is sufficiently low *and the reason that risk is low is because one takes the other to have goodwill toward one's self or one's valued goods;*

because one thinks another will be responsive to the fact that one is counting on him; or because one takes a participant stance toward the person one is interacting with. This counter-consideration does not hit its mark, however. Each of the three approaches I explain below is fundamentally different from understanding trust as a matter of risk-assessment. They do not understand trust to be the outcome of an assessment that the risk of counting on another is acceptable. Rather they understand trust to be a response to some trait of a given trustee or trustees despite the risk involved. On these approaches, trust is a response to a trustee having goodwill toward one, being trust-responsive, or being a morally responsible participant. That a trustee has one or more of those traits may in fact lessen the risk of counting on him. But on the approaches I will consider, that assessment and acceptance of risk is not the foundation for trust. A truster's relation to a trustee is, instead, the foundation of trust.

Section 4.2: Will-Based Approaches to Trust

Like risk-assessment approaches, will-based approaches are able to account for a significant amount of trust phenomena. But they are also able to account for at least some of the difference between trust and mere reliance. The main proponent of the will-based approach is Baier. I begin by outlining her view of trust as reliance on another's goodwill and competence before explaining Jones' critique of Baier and Jones' own will-based approach.

In her classic analysis of trust, Baier characterises trust as entrusting goods to another's discretionary care while relying specifically on their goodwill *and* competence (Baier, 1986, p. 234). In most of the places where Baier explicitly states her view she does not talk about competence. For instance, the first time she presents her view Baier describes trust as, "...reliance on their goodwill toward one, as distinct from the dependable habits, or only on their dependably exhibited fear, anger, or other motives compatible with ill will toward one, or on motives not directed on one at all" (Baier, 1986, p. 234). But later in the same paper Baier describes incompetence as one way that trust can be let down:

> One way in which trusted persons can fail to act as they were trusted to is by taking on the care of more than they were entrusted with—the babysitter who decides that the nursery would be improved if painted purple and sets to work to transform it, will have acted, as a babysitter, in an untrustworthy way, however great his goodwill. (Baier, 1986, p. 236)

I take this statement of Baier's to show that she thinks trustee competence as well as goodwill does matter to trust. Baier makes this explicit in a later paper where she writes that trust has a

> ...belief component, belief in the trusted ones' goodwill *and competence*, which then grounds the willingness to be or remain within their power in a way the distrustful are not, and to give the trusted discretionary powers in matters of concern to us. (Baier, 1994, p. 132. Emphasis added.)

I do not think that this later statement from Baier conflicts with her earlier work where the role of competence is less explicit. So while I refer to Baier's view primarily as a will-based view, I take her view of trust to include reliance on another's goodwill and competence. That said, I will focus on the goodwill component in Baier's account. This element distinguishes her account most substantively from other views and it is the feature that enables her view to account for the difference between trust and mere reliance. To understand the goodwill component of Baier's view, her analysis of trust as entrusting goods to another's discretionary care must first be explained.

Baier analyses trust on a three-place entrusting model, where *A* trusts *B* with valued thing *C* (Baier, 1986, p. 236). Baier provides two main reasons for thinking that trust should be understood in this way. First, she points out that individuals are unable to care for all that matters to them on their own (Baier, 1986, p. 231). In this Baier has pointed to a significant feature of human social life. Simply put, in our finitude we can care for more together than we can by ourselves.

> Given that I cannot myself guard my stamp collection at all times, nor take my rubber tree with me on my travels, the custody of these things that matter to me must often be transferred to others, presumably to others I trust. (Baier, 1986, p. 231)

Second, Baier analyses trust on an entrusting model to account for the fact that, although we need to trust each other with things, we don't trust everyone with everything. Rather, we tend to trust certain others with certain things. We do not trust everyone or even some people, with everything (Baier, 1986, p. 236).[13]

The need to leave at least some things in the care of others explains the practical vulnerability involved in trust. To the extent that we count on others to care for things that matter to us, we are vulnerable to them damaging those things or caring poorly for them. But Baier goes further, explaining that when we trust we leave things in the *discretionary* care of others. We leave others with room to care for us and/or our goods as they see fit and without checking up on them.

The discretionary element introduces what Baier calls "special vulnerability" (Baier, 1986, p. 239). Baier's concept of special vulnerability is different from the moral vulnerability I identified in section 3. Moral vulnerability involves risk of attitudes expressing ill will or indifference. In contrast, the vulnerability Baier has in mind is "special" in that it is *additional* vulnerability to practical harm that comes with giving trustees the benefit of the doubt and leaving them room to care for our valued goods as they see fit.

> If part of what the truster entrusts to the trusted are discretionary powers, then the truster risks abuse of those and the successful disguise of such abuse. The special vulnerability which trust involves is vulnerability to not yet noticed harm, or to disguised ill will. (Baier, 1986, p. 239)

On Baier's view the discretionary element, and the special vulnerability it establishes, are interconnected with the element of trustee goodwill at the centre of her analysis of trust.

[13] Hardin has also noted this context-specificity of trust. "...I might ordinarily trust you with even the most damaging gossip but not with the price of today's lunch (you always—conveniently?—forget such debts), while I would trust another with the price of lunch but not with any gossip. I might trust you with respect to X but not with respect to ten times X. Some few people I might trust with almost anything, many others with almost nothing" (Hardin, 2002, p. 9).

When we leave things in the discretionary care of another and become especially vulnerable to them, *and if we trust them*, then we will rely on their goodwill towards us and/or our goods. Likewise, if we take another to have goodwill towards us we will be more likely to leave things in their discretionary care thus making ourselves especially vulnerable. It is this element of reliance on another's goodwill that Baier thinks distinguishes trust from mere reliance. This is because it is possible to entrust goods to another's discretionary care without relying on her goodwill. We may be especially vulnerable to someone but merely rely on her fear or habits. Consider an alternate case of the Victorian worker example. Instead of an underprivileged worker, a rich and influential member of the upper class may shop at the markets and purchase products from the venders that he too knows have a tendency to tamper with their goods. This powerful member of society may rely on the market vendor to sell him unadulterated food because he knows she values his patronage highly. He relies on the vendor's fear of losing a valued customer. On Baier's view, the powerful member of society may rely on the vendor's *fear* of losing business but that would not be trusting her.

> What is the difference between trusting others and merely relying on them? It seems to be reliance on their good will toward one, as distinct from their dependable habits, or only on their dependably exhibited fear, anger, or other motives compatible with ill will toward one, or on motives not directed on one at all. (Baier, 1986, p. 234)

Baier's will-based approach has not gone unchallenged. Jones develops a substantial critique of Baier's view in her 1996 article "Trust as an Affective Attitude." There Jones argues against understanding trust on an entrusting model and offers her own view that trust involves an affective attitude of optimism about another's goodwill and competence.

Jones accepts that trust involves goodwill but she argues that trust should not be understood on an entrusting model because it is unable to account for the three characteristics of trust I identified in section 3: trust and distrust are contrary but not contradictory; trust cannot be willed; and trust gives rise to beliefs that are abnormally resistant to evidence. First, trust and distrust are contrary – one cannot trust and distrust someone at the same time and in the same domain. But they are not contradictory. Between trusting you and distrusting you, there is room for a third set of stances I can take toward you. I may merely rely on you without trusting or distrusting you. I may merely rely without trusting. Or I may just relate to you without relying or counting on you all together. And I may do so not because I necessarily distrust you. I may be merely not counting on you. I will discuss distrust further in *Chapter 5* when explicating the damage which betrayal can do to trust. For now I follow Jones in holding that distrust is contrary but not contradictory to trust. In contrast to this characteristic of trust, one either entrusts or does not entrust; there is no neutral stance. This is because entrusting is an action which either occurs or does not occur. For example, the separated co-parents either do leave their child in the other's care or they do not. There is no middle stance where one parent is neither entrusting nor failing to entrust the child to the other's care. On the entrusting view then, trusting (entrusting) and distrusting (not entrusting) are not only contrary but also contradictory.

Jones does note that entrusting models might meet her objection if trusting were to be seen as a positive handing over of something, distrusting as a positive refusal to hand over and the neutral stance, in between trust and distrust, as neither handing over nor withholding but instead taking no special precautions at all. Jones does not think such a fix would work in all instances of trust, however.

> There are three stances to take toward, say, you and the family silver: I may lend it to you, lock it up when I know you are visiting, or take no special precautions over it. It is less clear that there are three stances to take toward one's own self when walking down the street, and so it's not clear that this reply is fully satisfying. (K. Jones, 1996, p. 18)

I accept Jones' critique here. Understanding trust as entrusting fits some, but not all, trust phenomena.

Second, an entrusting model is unable to account for the fact that trust cannot be willed. This point of Jones' critique is straightforward. We can cultivate our trust by deciding that we have sufficient reason to trust another – as might occur in the case of developmental trust, which can involve significant doubts about whether the trustee will uphold our trust – but we cannot simply choose to trust. But in contrast, we can choose to entrust goods to another. As Jones writes, "Entrusting is an action and actions are, paradigmatically, things that can be willed" (K. Jones, 1996, p. 18). It is possible to make oneself entrust something to another's care but, while it is possible to cultivate one's trust, it is not possible to simply make oneself trust another.

Third, entrusting models are unable to account for the fact that trust can give rise to beliefs that resist relevant evidence. For example, if *A* and *B* are in a relationship characterised by trust, they will tend to interpret each other's behaviour in the light of that trust. Even if *B* acts in a way that raises suspicions of untrustworthiness among others, *A*'s trusting beliefs about *B* will be resistant to that evidence. Put another way, *A* may have "trust goggles" on. In contrast, Jones points out that beliefs providing reason to entrust something to another are not themselves resistant to evidence. Further, they do not give rise to beliefs that are resistant in that way either. For example, if I believe that you are a competent driver I may entrust you with my car. If you crash my car, I will most likely begin to doubt your driving skills. My belief about your driving competency is not resistant to relevant evidence. But on Baier's account we do not just entrust because we think another is relevantly competent. We entrust valued goods to others because we believe them to be competent *and have good will toward us*. And if I entrust you with my car because I believe you to be a competent driver and have good toward me, when you crash I may be slow to think you did so on purpose. I may begin to doubt your driving skills but continue believing that you have goodwill toward me. And so it seems that Baier's entrusting view *could* account for trust involving beliefs that are abnormally resistant to evidence. But Jones disagrees. She holds that entrusting models cannot account for trust's resistance to evidence, "...even when such a model is fully spelt out so that trust is entrusting on the basis of a belief that the other has goodwill" (K. Jones, 1996, p. 19). Jones points out that the seeming resistance to evidence accounted for by Baier's view could be explained by the fact that evidence for thinking another has goodwill is also evidence for believing they would not do something against us. Your kindness to me supports my belief that you have goodwill toward me. But it also supports my belief that you would not have crashed my car on purpose. I believe you are innocent, not because I believe you have goodwill toward me, but because I believe you are kind. But there is nothing about my beliefs about you on their own that should make them abnormally resistant to evidence (K. Jones, 1996, pp. 19-20).

I accept Jones' three criticisms and add to them the more general point that it is possible to leave valued goods – to entrust – while merely relying. For example, if I live in a small town with only one automobile mechanic, and she is known for doing an exceptionally poor job of fixing the type of car that I have, I may be in a position where I need to entrust the mechanic with my car though I do not trust her.

Rather than understanding trust on an entrusting model, Jones argues that trust is best understood as a type of affective attitude about another's goodwill and competence. Specifically, she explains trust as involving an affective attitude of optimism about another's goodwill and competence (K. Jones, 1996, p. 5). On her view, this attitude of optimism grounds an expectation that another will be directly and favourably moved by the thought that you are counting on them (K. Jones, 1996, pp. 5-6). I will return to consider this expectation in section 4.4 when I explain trust-responsiveness approaches to trust.

In talking about trust as involving an attitude of "optimism" Jones does not mean the general tendency to "look on the bright side" of things but rather a quasi-perceptual way of seeing another's goodwill and competence (K. Jones, 1996, p. 6). The type of quasi-perceptual attitude Jones thinks is involved in trust is similar to Ronald de Sousa's account of emotions as ways of seeing. On de Sousa's account emotions are perceptual. They have a *mind-to-world* direction of fit to the extent that they, "tell us things about the real world" (de Sousa, 1987, p. 203). But on de Sousa's account emotions are not merely perceptual. They also have a *world-to-mind* direction of fit as they draw our attention toward some things in the world and away from others. "Emotions make certain features of situations or arguments more prominent, giving them a weight in our experience that they would have lacked in the absence of emotion" (de Sousa, 2010). Emotions are then *quasi*-perceptual because they shape how we see events by highlighting features of our experience and drawing our attention to them. So too, the attitude of optimism Jones has in mind is a distinctive way of seeing others and their actions (K. Jones, 1996, p. 11). It directs its subject toward information consistent with thinking the other has goodwill toward them and away from contrary evidence.

For Jones, a truster's optimism is not global but rather qualified by the domain over which it extends. Individuals need not be optimistic to the same extent with every person or in different situations and they need not have the same level of optimism about one person in all areas of their relationship with that person. For example, I trust my sister more than I trust some of my friends. But I do not trust her equally with everything or in all situations. I may trust her to care for my dog while I am on holiday but I may not trust her to remember our mother's birthday. The *type* of attitude present – optimism – is the same in these situations but the level of my optimism – i.e. what I expect from another – differs. On Jones' account the attitude of optimism involved in trust leads to an expectation that another will be favourably and directly moved by the thought that one is counting on them (K. Jones, 1996, pp. 5-6). Again, I will explain this aspect of Jones' view in section 4.4 when discussing trust-responsiveness.

Will-based approaches have several benefits: they can account for the presence of the moral vulnerability trusters experience; they can explain the difference between trust and some types of mere reliance; and they can account for phenomena where a trustee's will plays a significant role. But they cannot account for the difference between trust and *all* types of mere reliance, and they fail to account for some trust phenomena.

Will-based approaches are able to explain moral vulnerability insofar as reliance and/or optimism about another's goodwill can make one susceptible to another's ill will or indifference toward one. To use Baier's phrase, when one relies on another's goodwill one is, "necessarily vulnerable to the limits of that goodwill" (Baier, 1986, p. 235). The trustee is in a position where she may take advantage of the truster's reliance on her goodwill. She may act out of ill will and violate the trust placed in her while keeping up appearances of having goodwill toward the truster. And on Jones' will-based approach, a truster is vulnerable to trickery or deception when his optimism about another's goodwill is resistant to contrary evidence. The distinctive way that a truster sees another is as one who has goodwill toward him. In reality the trustee may not have goodwill toward the truster. But the truster may be slow to perceive that fact due to the optimistic, "trust-goggles" he is wearing, leaving him vulnerable to moral harm.

Will-based approaches can distinguish trust from mere reliance on objects. This is because, as Jones has pointed out, if trust is reliance on another's goodwill you cannot trust something that does not have a will (K. Jones, 1996, p. 14). For example, I might rely on my car to get me to and from work each day without relying on its goodwill toward me – it does not even have a will let alone a good one toward me.

But within relations between persons, will-based approaches are less successful at accounting for the difference between trust and mere reliance. Holton has pointed out that a confidence trickster could count on a mark's goodwill toward him without trusting (Holton, 1994, p. 65). Rather than trusting the mark, the trickster is manipulating the mark and her goodwill. A similar critique can be made against Jones' view. It is possible to be optimistic about another's goodwill and competence without counting on them at all and it is possible to have such optimism while merely relying on them rather than trusting them. Again Holton's trickster is relevant here. Whether the trickster's stance toward the mark is articulated in terms of having a belief or an attitude of optimism about her having goodwill toward him, the trickster may still rely on the mark's having goodwill despite not trusting her. I will revisit Holton's example of the trickster and the mark when explaining the participant stance approach to trust below. That approach goes further than the will-based approach in explaining why Holton's trickster is merely relying on the mark rather than trusting her: because the trickster treats the mark as an object to be manipulated rather than taking a participant stance toward her.

In addition to failing to account for all of the difference between trust and mere reliance, will-based approaches fail to account for some types of trust. They *can* account for trust between strangers and intimates, but they have difficulty accounting for trust between enemies. For example, Baier's view *can* account for the trust between strangers involved in the Peruvian artist example. The artist might have left his stall in the patron's care because he thought that the patron would not be inclined to harm him or his stall. And in leaving the stall in the patron's care without checking up on the patron, the artist may have had little option but to rely on the patron's goodwill toward him and/or his work.

In a slightly different way, Baier's view can also account for the trust between separated co-parents. In that relationship, if the parents harbour animosity toward each other they each may have little reason to expect the other to have much goodwill towards them, but they may still trust because they take the other to have goodwill toward their shared child. So while the separated co-parents do not rely on each other's goodwill *toward them*, their relationship can still involve reliance on another’s goodwill.[14]

But Baier’s will-based approach is unable to account for some types of trust between enemies. For example, in the trust between enemies occurring in The Christmas Truce there does seem to be some goodwill present between the enemy sides, if only for one day. This goodwill was signified in the gift giving near the start of the truce. But as Holton points out not all trust between enemies will involve that element of goodwill. “Our enemies’ restraint might be grudging, driven simply by a belief that it is wrong to shoot someone who has surrendered” (Holton, 1994, p. 65). As I said when outlining the Christmas Truce example, Baier herself acknowledges that trust can occur between enemies (Baier, 1986, pp. 233-234). So she must think there is a way to understand interaction between enemies as reliance on the other’s goodwill. But I do not think Baier’s view can account for the situation Holton has identified. It is unclear that any sort of goodwill toward another is present when one restrains herself from harming an enemy just because she believes it is wrong to shoot someone who has surrendered. Such interaction between enemies may involve reliance on another’s moral code or personal convictions, but I do not think it involves reliance on another’s goodwill toward oneself.

In contrast, if Jones’ quasi-perceptual account is accepted, trust between enemies can be explained by holding that their optimism about each other’s goodwill might just be resistant to relevant evidence counting against their trust. The soldiers involved in The Christmas Truce might have been optimistic about each other’s goodwill despite the presence of evidence to the contrary – that they are engaged in a war with those they are counting on. Still, Jones’ view is not without its challenges. It is possible to trust despite not being optimistic about another’s goodwill toward you. Instances of developmental trust may involve such interaction. Recall the example of the parent who entrusts her teenage son with

[14] Holton also notes this possible interpretation of will-based approaches in his critique (Holton, 1994, p. 65).

the family car. If the son is especially rebellious, the parent may not be very optimistic about his goodwill or relevant competence. Yet, she may still trust him with the car in the hope of developing his ability to be trusted.

I have argued that, unlike risk-assessment approaches, will-based approaches are able to account for at least some of the difference between trust and mere reliance but that they are limited in their ability to account for some phenomena. I now turn to consider an approach to trust – which I will call the *participant stance approach* – that is able to account for more phenomena than goodwill-approaches but which is still unable to account for all of the difference between trust and mere reliance.

Section 4.3: Participant Stance Approaches to Trust

Noting the limitations of the will-based approach in accounting for some types of trust phenomena, Holton and others[15] have sought an alternative explanation for why trust is distinct from mere reliance. Specifically, their search is directed by a premise that I introduced in the introduction to this thesis: that trust is vulnerable to betrayal but mere reliance is not.[16] As I noted in the introduction of this thesis, this premise cannot do substantive work until an account of betrayal has been articulated; something that Holton and others have left undone. In *Chapter 4* I provide such an account of betrayal. But in the meantime I will accept that it is prime facie true that mere reliance can only be let down or disappointed but trust can be disappointed, let down *and* betrayed.

Holton and those following him have sought to explain the distinct vulnerability to

[15] McGeer, (McGeer, 2002, p. 33); Pamela Hieronymi, (Hieronymi, 2008, p. 215); Stephen Wright, (Wright, 2009).

[16] Philosophers who distinguish trust from mere reliance in terms of vulnerability to betrayal include Baier, (Baier, 1986, p. 235); Walker, (Walker, 2006, p. 85); McGeer, (McGeer, 2002, p. 33); Hieronymi, (Hieronymi, 2008, p. 215); Wright, (Wright, 2009); McLeod, (McLeod, 2011); and Holton, (Holton, 1994, p. 66) Of these philosophers all but Baier and Walker use trust's vulnerability to betrayal to motivate a participant stance approach to trust.

betrayal involved in trust by pointing out that persons only feel betrayed when they are let down by other persons rather than by objects. As Holton writes,

> In cases where we trust and are let down, we do not just feel disappointed, as we would if a machine let us down. We feel betrayed. [...] The central point is rather that betrayal is one of those attitudes that Strawson calls the reactive attitudes. These are attitudes that we normally take only towards people. (Holton, 1994, p. 66)

Others have followed Holton in holding that trust is reliance with a stance that is only taken toward persons.[17] McGeer puts the central claim of the participant stance approach well in the following remark about trust:

> It is a state, rather, that encourages us to invest trust in them, that encourages us to see them as capable of being responsive to our trust. Hence it is an attitude we take towards the character of their agency—in part, I will argue, by taking the same attitude towards our own. (McGeer, 2008, p. 242)

Philosophers who hold that trust involves an attitude that is only taken toward persons have looked to Strawson's theory of participant reactive attitudes to explain the vulnerability to betrayal – or at least the vulnerability to *feeling* betrayed – involved in trust (Holton, 1994, pp. 66-67)[18]. The idea is that feeling betrayed is a participant reactive attitude so trust must involve something like taking a participant stance toward one you are relying on.

After briefly explaining Strawsonian participant reactive attitudes I will explain the participant stance approach towards trust argued for by Holton. I will then explain why that approach is able to account for more trust phenomena than the will-based approach but is too

[17] See: Walker, (Walker, 2006, p. 85); McGeer, (McGeer, 2002, p. 33); Hieronymi, (Hieronymi, 2008, p. 215); and Wright, (Wright, 2009).

[18] For other views connecting trust to Strawsonian participant reactive attitudes see: Hieronymi, (Hieronymi, 2008, p. 215); McGeer, (McGeer, 2008, p. 33); and Wright, (Wright, 2009). In contrast, Walker has argued that at the core of Strawsonian participant reactive attitudes is a stance of holding one responsible. Rather than just looking to reactive attitudes, Walker argues that trusters can experience negative reactive attitudes when they are let down while those merely relying do not, because trusting involves having a normative expectation that others can be held responsible to meet (Walker, 2006, pp. 79-81). I accept Walker's view that Strawson's reactive attitudes are about responsibility (Strawson, 1974, p. 6). I also take her point that negative reactive attitudes involved in trust can be explained by the presence of normative expectations. But I do not think her critique of Holton alters my analysis of participant stance approaches. Her view has the same outcome: trust involves taking something like a participant stance because one has a normative expectation of another. Further, Walker herself describes her account as capturing the participant stance involved in trust (Walker, 2006, p. 80). So I will consider Walker's normative expectation view to be another variation of the participant stance approach.

broad to account for the difference between trust and mere reliance.

In *Freedom and Resentment* Strawson explains what he calls “participant reactive attitudes” (Strawson, 1974, p. 10). These are attitudes that persons take toward others participating in given relationships with them. Strawson contrasts such attitudes with those taken toward objects and other persons behaving in ways not taken to express states of will or intentions toward oneself (Strawson, 1974, pp. 5-10). For example if a book falls on my foot I may react with expressions of pain but, unless I take an anthropomorphic view toward the book, I will not experience, or express, resentment toward the book. But if a person comes and intentionally stomps on my toe I will most likely feel and express both pain and resentment.

I take the participant stance to be a disposition to respond with certain attitudes should we take another’s behaviour to express goodwill, ill will or indifference toward us. This interpretation is supported in the following statement by Strawson: “What I have called the participant reactive attitudes are essentially natural human reactions to the good or ill will or indifference of others towards us, as displayed in their attitudes and actions” (Strawson, 1974, p. 10).

McGeer develops the implications of Strawson’s view further, explaining that the reactive attitudes involved in taking a participant stance toward someone are best explained as “co-reactive” (McGeer, 2011, p. 8). That is, the attitudes in question are reciprocal. They respond to states of another’s will but also express states of the subject’s will and in doing so elicit further reactive attitudes from the other. For example, if you intentionally stomp on my toe, the ill will that your action expresses toward me elicits a reaction of resentment in me. And when I express that resentment to you, my resentment calls for a further response of apology from you which then, if you do apologise, elicits a response of forgiveness from me.

This “calling for response” highlights what McGeer calls the “forward-looking dimension” of participant reactive attitudes (McGeer, 2011, pp. 8-9). She writes,

> ...reactive attitudes are backward-looking responses to the actions and attitudes of others, to be sure. But, more importantly, they have a forward-looking dimension, serving to elicit some further reactive response from the individuals to whom they’re directed. (McGeer, 2011, pp. 8-9)

I accept McGeer’s interpretation of Strawson’s view and take it to explain the participant attitude as involving seeing another as one toward whom we are disposed to experience reactive attitudes should her behaviour express goodwill, ill will or indifference toward us. It also involves seeing oneself as responsible for responding to the other’s reactions. Given this explanation, the participant attitude is not just a stance one takes toward another, or others; it is a stance one takes toward another, or others, *and* toward oneself.

According to Strawson, the participant stance is not taken toward all persons or even toward any person all the time (Strawson, 1974, pp. 7-8). There are some human beings, he thinks, toward whom it would be inappropriate for us to take a full-blown participant stance. These individuals are those without the requisite agency for their behaviour to be taken to express states of their will toward others. Strawson identifies individuals that are “warped or deranged, neurotic or just a child” (Strawson, 1974, p. 8) in this category. Because such beings are either psychologically abnormal or morally underdeveloped, they are exempt from having their actions taken to express states of goodwill, ill will or indifference toward others. But there are also persons whose behaviour would normally be taken to express such states but who are momentarily excluded from being seen as such. They are excluded for a time, but are not normally exempt from responsibility for their behaviour. This is the case with persons whose behaviour is taken to have been done under duress or in, as Strawson puts it, “abnormal stresses” (Strawson, 1974, p. 8). So to take a participant attitude toward someone is to see them as having the capacity to act in ways expressing states of good or ill will, presuming that no excusing conditions apply.

Based on Strawson's participant attitude, proponents of the participant stance approach to trust understand trust as *reliance that involves a readiness to experience participant reactive attitudes toward the one that is being relied upon.* This interpretation of the approach is evident in Holton's claim that "When you trust someone to do something, you rely on them to do it, and you regard that reliance in a certain way: you have a readiness to feel betrayal should it be disappointed, and gratitude should it be upheld" (Holton, 1994, p. 67).[19]

By identifying a participant stance in trust, I think Holton and others make explicit and clarify something which is implicit in will-based approaches: that the "object" of one's trust is the kind of being that is capable of goodwill, ill will and indifference. This claim is implied in the ability of Baier's will-based approach to distinguish between trust and mere reliance on objects. As I explained in section 4.2, on Baier's account I can merely rely on my car but not trust it because my car does not have a will that I can rely on. As Jones has pointed out, "One can only trust things that have wills, since only things with wills can have goodwills..." (K. Jones, 1996, p. 14). The claim that trustees are the kinds of things who can have and express goodwill – and ill will and indifference – is what participant stance approaches develop. In developing that claim, participant stance approaches clarify the background assumption in will-based approaches that trust involves counting on one who is capable of having goodwill toward you.

Understanding trust as involving a participant stance is able to account for the difference between trust and some instances of mere reliance. For example, it can explain why the trickster discussed earlier is not trusting her mark despite relying on the mark's goodwill. The reason is that the trickster treats the mark as an object to be manipulated rather than as a participant (Holton, 1994, p. 67). The trickster is not engaged in a reciprocal

[19] Hieronymi explicitly follows Holton in this. "On Holton's account, trust is *reliance from the participant stance*. Thus far, I agree" (Hieronymi, 2008, p. 216 Emphasis is Hieronymi's).

relationship with the mark but, rather, uses her. With regard to reliance relationships like the example of the trickster – where a truster is not taking a participant stance toward the object of his reliance – the participant stance approach is able to account for the difference between trust and mere reliance.

The participant stance approach is also able to account for the examples I presented in section 2. Each of those examples involves trust between persons. As such, each truster and trustee could be described as being disposed to feel reactive attitudes in response to the other's actions toward him. For one example, consider the "trust fall" again. If the blind-folded person lets herself fall and the others try to catch her but fail to do so because they are too weak, the blind-folded person may be disappointed but she will probably not feel resentment toward her teammates. But if the blind-folded person falls and the teammates do not even try to catch her, she may take their omission as expressing ill will or indifference toward her and feel not just practically but also morally harmed. In response she may feel resentment toward her teammates and possibly even feel that they have betrayed her.

The participant stance approach is also able to account for the different features and types of trust I identified in section 3. For example, persons can take a participant stance toward strangers, acquaintances or intimates. And it can be taken towards others in one-off interactions or in iterated relationships. Further, there is no reason to think that the participant stance is inconsistent with the acceptance of vulnerability involved in trust. It only specifies that trusters accept vulnerability to participants while those merely relying may also accept vulnerability to objects and individuals excluded or exempted from responsibility for their behaviour.

In addition, the participant stance approach can explain the moral vulnerability involved in trust. As I have shown, risk-assessment approaches, will-based approaches and trust-responsiveness approaches can allow for the fact that trust involves moral vulnerability.

But those approaches do not in themselves explain *why* trusters would be vulnerable to moral harm. On a participant stance approach, moral vulnerability is explained by virtue of the object of one’s trust being a person whose attitudes and intentions matter to us. As a participant rather than an object, the behaviour of trustees can be taken to express states of their will. And so, should the trustee let down the truster, the trustee’s actions might be taken to express not just a lack of ability to do what the other was counting on her for; it might also be taken to express ill will or indifference toward the truster.

The participant stance approach is able to account for some of the difference between trust and mere reliance and for the above trust phenomena. But it is too general to account for all of the difference between trust and mere reliance. It is possible to take a participant stance toward someone while merely relying on, rather than trusting, him. Consider again the Victorian vendor who sells adulterated goods. In mixing dirt into cocoa and then selling it to the worker, the vender shows indifference toward him and other workers. The worker might take a participant stance toward the vendor and rely on her without trusting. That is, he might be prepared to feel resentment towards the vendor and rely on her for food but not think that she is someone to be trusted. Unlike the trickster in Holton’s example who merely uses the mark, the worker would be taking a participant stance toward the vendor. Unlike the trickster, he would be engaged in a reciprocal relationship with her and be responsive to her expression of indifference toward him. But because of that attitude of indifference the worker does not trust the vendor.

Unlike the participant stance approach, the trust-responsiveness approach to trust, which I consider next, *is* able to account for the difference between trust and mere reliance in instances like the Victorian worker example. Further, as I will explain, trust-responsiveness views can be taken to assume a type of participant stance. So, as I will argue, among other things, trust-responsiveness approaches can explain the difference between trust involving a participant stance and merely relying on another while taking a participant stance toward them. But such approaches are also not without their challenges; they are still unable to account for some of the difference between trust and mere reliance and for some trust phenomena.

Section 4.4: Trust-Responsiveness Approaches to Trust

In this section I revisit Jones' 1996 account of trust as an affective attitude, and I explain the expectation of trust-responsiveness that she takes to stem from the attitude of optimism about another's goodwill and competence. Having explained the trust-responsiveness aspect of Jones' account, I explain why I think that her account assumes a participant stance. I then argue that Jones' view is able to distinguish between trust and mere reliance. But because it stems from a will-based approach, it is still unable to account for phenomena where trustee will does not play a significant role. But not all trust-responsiveness approaches are limited in that way. To show this I explain a trust-responsiveness approach from Pettit which is not dependent on a goodwill approach. But as with the other approaches I have considered, Pettit's view is still unable to account for some trust phenomena.

As mentioned in section 4.2, Jones holds that the attitude of optimism about another's goodwill and competence grounds an expectation that the other will be favourably responsive to the fact that you are counting on them. Because trusters are optimistic about another's goodwill toward them, they expect that the trustee will be moved by the fact that the truster is counting on him. But, according to Jones, trusters do not just expect trustees to be favourably trust-responsive but to be *directly* so. This means that the fact that I am counting on you does not just motivate you to respond favourably to me *as a means to some other end* or *for some other reason*. Rather, it motivates you to be moved just because I am counting on you. The qualification that trustees are "directly" moved by the thought that they are being counted on helps to distinguish some types of trust from mere reliance (K. Jones, 1996, p. 9). For example, consider the “trust fall” exercise. If the blind-folded person expects the teammates to catch her just because they are afraid of her anger towards them should they do otherwise, something of trust is lost. It seems that rather, in the trust exercise, the blind-folded person falls expecting to be caught just because she is counting on the others. Perhaps this is why the “trust fall” exercise is used for building cooperation and relationships among teammates who may not know much about each other or have trouble cooperating – it requires no further knowledge about those involved other than that they are counting on you to catch them and you will need to do the same when it is your turn to be blind-folded.

The expectation of trust-responsiveness involved in Jones’ account need not involve an expectation that trustees will always carry out favourable actions toward those counting on them. On the contrary, *A*’s counting on *B* to do *X* may give *B* reason to do *X* while having overriding reasons not to do *X*. And even when motivation to respond favourably to *A*’s dependence prevails, *B* may have different options for how to favourably respond (K. Jones, 1996, p. 8). Trustees must then exercise discretion as they consider whether to respond to another’s trust and, if so, how to respond. For instance, in the Peruvian Artist example, the patron who stays behind to work the art stall while the artist and the other buyer make the delivery must exercise discretion in how far to go in caring for the selling of the artist’s work. The patron will at least be expected to guard against thieves at the stall and to care for the money earned if any sales are made while the artist is away. But the patron would probably be going too far if she were to rearrange the placement of the artist’s work in the stall or sell the work for a lower price than the artist had marked just to make a sale.

Jones’ understanding of trust as involving an expectation of trust-responsiveness is able to account for the difference between trust and mere reliance on inanimate objects. This is because such objects cannot be expected to do that which we are counting on them for, just by virtue of the thought that we are counting on them to do so. Thus, her account assumes that trustees have the capacity to be responsive to our counting on them. My car does not have that capacity and so I can rely on it but I cannot trust it. Further, I do not think the capacity to be responsive to the thought that another is counting on you is very different from having the capacity for Strawsonian participant status. This is because participant reactive attitudes are essentially responsive to the attitudes and intentions of others toward us (Strawson, 1974, p. 6). Instead, I think "being responsive to the thought that another is counting on you" is one type of responding to another's attitudes or intentions. In other words trust-responsiveness is a reactive attitude that responds specifically to another's attitude when accepting vulnerability to us. That attitude may be expressed explicitly to a trustee but it may also be manifest in the truster's behaviour. Either way, trust-responsiveness is a reaction to the expression of an attitude to accept vulnerability and count on another. So part of the reason I rely on my car, but do not trust it, is that it is not the kind of thing that can participate in relationships with me and react to my expressed attitudes toward it.

Understanding trust as involving an expectation of trust-responsiveness can also account for the difference between trust and mere reliance on participants. As I pointed out when criticising the participant stance approach to trust, it is possible to rely on another and to take a participant stance toward that other without trusting them. I used the example of the Victorian worker relying on the market vendor to illustrate this possibility. But understanding trust on a trust-responsiveness approach accounts for the difference between trust and such reliance: the Victorian worker may take the vendor to be a participant but not expect her to be directly and favourably moved by the thought that he is counting on her. That is, he takes a participant stance toward the vendor but does not specifically expect her to be trust-responsive. On Jones' view, the worker does not expect the vendor to be trust-responsive because he does not think she has goodwill toward him, which fits the example. Still, if it is possible to rely on someone and expect them to be directly and favourably moved by the thought that I am counting on them without trusting them, then the trust-responsiveness approach is unable to account for *all* of the difference between trust and mere reliance. Such a counterexample can be found. Recall Holton's confidence trickster who relied on the mark's goodwill but did not trust him. The confidence trickster could expect the mark to be moved by the fact that she is counting on him and yet not trust the mark. She merely relies on his responsiveness. So, like the risk-assessment approach, goodwill approach and the participant stance approach, understanding trust in terms of trust-responsiveness is unable to account for all of the difference between trust and mere reliance.

Since it is grounded in trustee goodwill, Jones' trust-responsiveness view is limited in its ability to account for phenomena where trustee goodwill does not play a significant role. So while there might be some reason for enemies to be responsive to the fact that they are counting on each other, thinking the other has goodwill toward one is probably not operative in such an interaction. But it is possible to avoid this objection if a trust-responsiveness approach is not also a will-based approach. Pettit has taken such an approach.

While Pettit understands "trust" to refer to a variety of types of reliance he identifies a specifically trusting type of reliance in terms of trustee responsiveness to being counted on – what he calls, "interactive trusting reliance" (Pettit, 1995, p. 206). He arrives at that specific type of reliance through making a series of distinctions (Pettit, 1995, pp. 203-206). First, he narrows his focus by first distinguishing between reliance on natural phenomena and reliance on people. While he thinks it is appropriate to talk about reliance on natural phenomena as "trust" his concern is with reliance on people. Second, Pettit distinguishes between reliance on people to have certain skills or capacities and relying on them to act in specific ways. While Pettit does not acknowledge this, relying on someone to have certain skills or capacities will also mean relying on them to act in specific ways. But I take Pettit's point to be that it is possible to rely on someone to act in a specific way without relying on them to have certain skills or capacities. For example, if I rely on my flatmate not to eat all of the food in our fridge, it is not clear that I am relying on him to have any specific skills. But I am relying on him to act in a specific way. Third, within reliance on persons to act in a specific way, Pettit singles out placing oneself in the hands of another as distinct from being confident that people will be reliable under certain tests or just having a general confidence that people will act in certain ways. Fourth, within instances of placing oneself in the hands of another, Pettit distinguishes between what he calls "active" reliance on people (i.e. choosing to make yourself vulnerable to another) and doing so out of necessity or because you are being forced to do so. Fifth, he highlights reliance that is not just active but interactive. He explains interactive reliance as involving an expression of one's reliance to another on whom one is relying. Sixth, and lastly, Pettit highlights reliance that is not just interactive but interactive in a distinctively trusting way. This "interactive, trusting reliance" (Pettit, 1995, p. 206) involves an expectation that another will do *X* in a given way because, when made aware that the truster has made himself vulnerable to her and is relying on her to do something, she is moved to do what is expected of her.[20]

Unlike Jones' view, Pettit's trust-responsiveness view is not grounded in goodwill since persons can be *positively moved when made aware that a truster has made himself vulnerable to them* for various reasons. One of these reasons may be that the trustee has goodwill toward the one counting on her. But Pettit identifies a variety of reasons persons may be positively responsive. They may be responsive for the sake of being loyal; because they think it would be virtuous to do so; because they think doing so would be prudent; or

[20] Pettit is careful to say that he does not think *all* trust is interactive-trusting-reliance. Rather he just uses the trust/reliance distinction to point to one distinct type of trust. He writes, "Interactive, trusting reliance, as I have stressed, is not the only thing that we use the word 'trust' for. When I focus on such reliance, I do not mean to suggest that it has any monopoly claims on the name of 'trust.' And when I identify the conditions under which such reliance is present, I do not mean to present them as conditions in the analysis of the concept of trust" (Pettit, 1995, p. 207).

because they think doing so will make the truster, and third-parties, have a good opinion of them (Pettit, 1995, pp. 209, 215). By leaving open the reasons why persons might be trust-responsive, Pettit's view is able to account for more types of trust phenomena than Jones' trust-responsiveness approach. It can account for trust where goodwill does play a significant role and it can account for instances where persons are expected to be trust-responsive for other reasons.

But Pettit's view is not without its own challenges. I do not think it provides the best account of the trust between separated co-parents. This is because it is not clear that the parents' trust is based on an expectation that each will be responsive to the thought that the other has made himself or herself vulnerable to the other. Rather, I think it is likely that trust between them is based on an expectation that the other will be responsive to the vulnerability of their shared child. That amounts to a type of trust-responsiveness as long as the relationship we are identifying is between the child and one of the parents. But there is a relationship between the co-parents themselves that involves trust. And yet, I do not think the parents necessarily expect each other to be responsive to the fact that the other is counting on them. So while the trust-responsiveness approach is able to account for some of the difference between trust and mere reliance, and is able to account for some types of trust, it is not able to account for all types of trust. And that is the case even if the expectation of trust-responsiveness is not grounded in an element of goodwill.

In this section I have explained four main influential approaches to trust: trust as risk-assessment; trust as reliance on the goodwill of another toward you; trust as a participant stance toward another; and trust as an expectation of trust-responsiveness. I have argued that each approach has some benefits but is ultimately unsatisfactory. None of the approaches is completely able to account for trust's varied phenomena and its difference from reliance.

The limitations of the approaches I have considered need not nullify their usefulness for understanding some instances of trust. For example, not all trust is best understood as reliance on another’s goodwill, but goodwill is still important to many cases of trust. Goodwill may factor strongly (or not) in trust between unseparated co-parents, or potentially in the relationship between the Peruvian artist and his patrons. He may count on the patrons precisely because they seem to be nice people who would not have any reason for ill will toward him. Similar points can be made about risk-assessment, the participant stance and trust-responsiveness. Some trust will involve risk assessment and arise between strangers taking part in a one-off relationship. Other cases of trust may look more like instances of taking a participant stance toward someone who is known very well and shares an extended relationship with the truster. Still other cases of trust may best be explained as involving an expectation of trust-responsiveness, perhaps between acquaintances where a vulnerability to moral harm is very pronounced.

But while the approaches I have considered can be useful, a satisfactory understanding of the concept of trust as a whole is still needed. It is possible that the approaches I have considered simply point to the wrong conditions for trust. There may be a different type of assessment, belief, expectation or attitude that all trust phenomena could be found to involve. But I think this is unlikely. Because of the breadth and variety of the phenomena, I am pessimistic about identifying one set of necessary and sufficient conditions that all instances of trust will meet. Either some phenomena will remain unaccounted for or the concept will be made too inclusive, compromising trust’s distinctiveness from related concepts. Instead, I propose that we should reconsider what *type* of concept trust is.

Section 5: Trust as a Cluster Concept

Some concepts are structured around a set of necessary and sufficient conditions. For example, we might take the condition of “being something that breaks when it is struck” to be a necessary and sufficient condition for the concept of fragility.[21] Any instances of fragility – any fragile things – would, then, have the common feature of breaking when struck, or being likely to break if struck. But not all concepts are like this. Ludwig Wittgenstein has pointed out that some concepts do not have features common to all their instantiations. Instead, instances of such concepts involve various similarities and relationships.

> Consider for example the proceedings that we call ‘games’. I mean board-games, card-games, ball-games, Olympic games, and so on. What is common to them all?—Don’t say: ‘There *must* be something common, or they would not be called ‘games’’—but *look and see* whether there is anything common to all.—For if you look at them you will not see something that is common to *all*, but similarities, relationships, and a whole series of them at that. To repeat: don’t think, but look!... (Wittgenstein, 1958, p. 27. Emphasis is Wittgenstein's)

Wittgenstein refers to the similarities and relationships occurring in concepts without common features as “family resemblances” (Wittgenstein, 1958, p. 27). Like human families, particulars of this type of concept will resemble each other in different ways and each need not be similar to the others in exactly the same way. For example, one child in a family may have her father’s eyes and her mother’s nose while a sibling has the mother’s nose but not the father’s eyes. If I were to look at a photograph of the whole family, I might be able to tell that they were all related by their physical features, but each would have different similarities to individual family members. And there may not even be a single distinctive feature that each member of the family has. It is this kind of similarity and relation that Wittgenstein has in mind.

My analysis in section 4 essentially amounts to looking at trust phenomena to see if there are in fact any common features that all instances share. The outcome of my analysis points to trust being the type of concept described by Wittgenstein’s idea of family resemblance. We may postulate that all trust involves some type of risk-assessment, relying on another’ goodwill, expecting another to be trust-responsive or taking a participant stance

[21] I take this example from David Lewis (Lewis, 1997, p. 143).

toward another. But when I follow Wittgenstein’s direction and *look* at the phenomena, that is not what I see. As I have shown, instances of trust can have similarities but they can also have dissimilarities. I have not found any features that all trust phenomena share in common but that instances of related concepts do not share. For example, risking and being vulnerable are quite central to trust phenomena. It may be thought that they are the common features in instances of trust. But even if all trust involves vulnerability and risk, those features are not distinctive of trust. As I have tried to show, it is also possible to risk and be vulnerable while merely relying.

My suggestion then is that trust is a concept that is not amenable to analysis in terms of necessary and sufficient conditions. Rather I think it should be understood as the type of concept identified by Wittgenstein’s idea of family resemblance. I follow Stoljar in understanding such concepts as *cluster concepts* (Stoljar, 1995, p. 283). Cluster concepts involve a group of features that tend to cluster together but which are not necessarily common to all instances of a given concept. Some of these clustering features will be more central than others. To use Wittgenstein’s example of the concept “game”, *being fun* may be more common to games than there being *a winner and a loser*. But not all games will be fun – e.g. war games. And not all games will have a winner and a loser. To borrow an example from Wittgenstein, “In ball-games there is winning and losing; but when a child throws his ball at the wall and catches it again, this feature has disappeared” (Wittgenstein, 1958, p. 27). Neither feature is common to all games but if being fun turns out to be more common it can be understood as a more central feature in the cluster of “game.” There is bound to be a number of central features in any given cluster concept.

In analysing the concept “woman” as a cluster, Stoljar identifies particulars with a significant number of central features as exemplars or paradigm cases of a cluster (Stoljar, 1995, p. 284). For example, she identifies the following features of the concept “woman”:

female sex, phenomenological aspects of what it feels like to be a woman, specific gender roles, wearing typically female attire, being oppressed on the basis of one's sex, and self-attributes such as calling oneself a woman (Stoljar, 1995, p. 284). Particular individuals who have these features will be exemplars of the concept "woman."

The idea of exemplars of cluster concepts aids in understanding border-line cases of a concept. For a particular to count as an instance of a cluster it need only be sufficiently similar to an exemplar. To borrow Stoljar's example again, "individuals of indeterminate sex who have female lived experiences, social roles, and self-attributions of womanness will count as women" (Stoljar, 1995, p. 285).

When trust is understood as a cluster concept, elements that are more stable across trust phenomena can be considered more central features of its cluster. For instance, features clustering together consistently in the examples discussed in this chapter include counting on someone, vulnerability to practical harm, vulnerability to moral harm and accepting vulnerability. These features could, arguably, be considered central features of trust. In contrast, features that tend to vary across examples of trust may be less central, such as variation in degree of intimacy (i.e. counting on a stranger, acquaintance or intimate); variation in degree of duration of interaction (i.e. counting on another in a one-off interaction or in an iterated relationship); variation in the level of confidence involved in accepting vulnerability to a trustee (i.e. confidently accepting vulnerability or doing so despite the presence of doubts); and the minimization of risk, reliance on or optimism about trustee goodwill, a participant stance or trustee trust-responsiveness.

In a given interaction, any number of the above features may be salient but no set number of them need be present for it to count as trust. Interactions involving trust's central features will be exemplary cases. For example, the "trust fall" example will be an exemplary case as it involves counting on others, vulnerability to practical harm, vulnerability to moral harm and acceptance of vulnerability. That case may also involve relying on the goodwill of others toward oneself, expecting others to be trust-responsive and entrusting one's bodily safety into the care of others.

All instances of trust need not be exemplary cases though. For a given interaction or relationship to be considered an instance of trust it must just be sufficiently similar to an exemplary case like the "trust fall."[22] For example, I have explained that the separated co-parent example involves counting on another (in this case a former intimate), being vulnerable to another and accepting that vulnerability and entrusting a valued child to another's care. Since the separated co-parent case shares several elements with the exemplary "trust fall" case, it will count as an instance of trust.

Understanding trust as a cluster concept explains why trust language is often used to refer to interactions that involve only minimal features of trust. For example, consider talk of placing trust in inanimate objects. When I say I "trust my car to get me to work" the only salient feature doing much work is my vulnerability to the practical harm of the car breaking down. But my reliance on my car is not sufficiently similar to exemplary cases of trust. For instance, in the exemplary "trust fall" case the blind-folded participant lets herself fall, hopefully into the arms of her teammates, because she thinks it is in their interests to catch her; because they have goodwill toward her; or perhaps because she thinks they will be trust-responsive. But I do not rely on my car for such reasons. I may simply think the car was well built. So, on my view, my reliance on the car shares some features with instances of trust but does not count as an instance of trust itself.

At the beginning of section 2, I said that I would not be motivating my analysis of trust with examples of counting on natural phenomena. My reason for doing so was that it is unclear whether talk of "trust" in natural phenomena should be taken at face value. With my

[22] This idea that instances of trust are sufficiently similar to exemplary cases tracks Stoljar's explanation of the concept of 'woman' as a cluster concept (Stoljar, 1995, p. 284).

understanding of trust as a cluster concept in hand I can now explain why it is less clear that statements such as “I trust the sun to rise in the morning” identify instances of trust. Like talk about “trusting” inanimate objects, such language identifies situations that are not sufficiently similar to exemplary cases to count as instances of trust. When I “trust” the sun to rise, I may be *vulnerable* to the sun not rising – and that is a significant vulnerability! – but that is where the similarities with paradigm cases of trust end. I am not morally vulnerable to the sun nor do I rely on its goodwill; take a participant stance toward it; or think it will be trust-responsive. So, I think talk of “trust” in natural phenomena should also be understood metaphorically. The cluster concept account of trust explains why it is understandable that we might talk about “trust” with regard to inanimate objects or natural phenomena. But it also explains why we should disregard such linguistic usage when analysing the central features of the concept of trust.

Understanding trust as a cluster concept also explains border-line cases in interpersonal phenomena. For example, Baier has talked about the infant/parent relationship as a relationship of trust (Baier, 1986, p. 241). However, while the infant is an exemplar of vulnerability – since he must depend on others for everything – it is not clear that this relationship is appropriately characterised in terms of an infant *trusting* his parents. The relationship may be better understood as one of mere vulnerability, or mere reliance, rather than trust. Nevertheless understanding trust as a cluster concept provides an explanation of why our intuitions regarding such border-line cases may diverge: such cases involve some of the features of trust’s cluster but it is unclear whether they are sufficiently similar to exemplary cases or not.

An account of trust as a cluster concept does not face the same limitations as the influential approaches I considered in section 4. Because it does not attempt to identify necessary and sufficient conditions for trust, the cluster approach is not challenged by the variety across trust phenomena. It can also account for the difference between trust and mere reliance. On the cluster approach, instances of mere reliance will be cases of reliance that are not sufficiently similar to paradigm cases of trust. They may share some features with instances of trust, which explains why it is possible to interact with another because the risk of doing so is acceptable, rely on another's goodwill, expect trust-responsiveness, or take a participant stance toward another while merely relying. But instances of mere reliance will not share enough features with paradigm cases of trust to be sufficiently similar to them. So they will not be instances of trust themselves. Because the cluster approach can account for the variety in trust phenomena and for the difference between trust and mere reliance, I take it to provide a satisfactory understanding of the concept of trust. I now have a philosophical analysis of the concept of trust that I will be able to employ in analysing betrayal.

In *Chapter 2* I shift my focus from trust to trustworthiness. Both concepts will prove important in my analysis of betrayal in *Chapter 4* and in my explanation of the interplay between trust and betrayal in *Chapter 5*. Presumably, whatever betrayal is, part of what its victims experience is disappointment, and even violation, by someone they thought it was reasonable to trust. But whether that means victims are affected by an individual they took to be *trustworthy* or not needs to be explained. In *Chapter 2* I show that there is reason to think being trustworthy is not just the same as being someone whom another has reason to trust.

Chapter 2: Trustworthiness, Trustability, and Mere Reliability

Section 1: Introduction

When we trust others we risk them harming us and our goods. Because of that risk, it is important to know who we have good reason to trust – i.e. who it is *reasonable* to trust. A promising, but uninformative, answer is that trust is reasonable when placed in the trustworthy. This answer is promising because there is, putatively, a special relation between trust and trustworthiness. The trustworthy are, if nothing else, worthy of our trust. And it is at least prime facie true that we have good reason to place our trust in those who are worthy of it. But, this answer is uninformative unless we have a clear understanding of what it means to be trustworthy.

The aim of this chapter is to provide an explanation of trustworthiness. I develop two central arguments: that the concept of trustworthiness is best understood as indicating a kind of virtuous character; and that those who are trustworthy are not the only persons it is reasonable to trust. We can also have good reason to trust others despite not knowing much about their character. To distinguish the smaller class of those who are trustworthy from the larger class of those we have good reason to trust I introduce the concept of "trustability." On my account, being trustworthy is one way to be trustable, but it is not the only way. Persons can also be trustable – that is, it can be reasonable to trust them – because of the character of their wills toward us, because we expect them to be trust-responsive, or because they have various incentives to uphold our trust.

Two constraints will guide my analysis throughout this chapter. First, a satisfactory account of whom it is reasonable to trust should reflect the variety in trust phenomena. For example, it seems plausible that good reasons for trusting intimates would differ from those supporting trust in strangers. For one thing, trusters have much more knowledge of intimates than strangers. And yet, as I showed in C*hapter 1*, trust can be placed in strangers. I think it is unlikely that *all* trust in strangers is unreasonable. Rather, it is likely that there are good reasons for trusting strangers; but they will probably differ from reasons for trusting intimates. A satisfactory understanding of whom it is reasonable to trust should account for such variation.

Second, a satisfactory account of trustworthiness should be able to distinguish it from related concepts without failing to account for relevant phenomena. As with trust and mere *reliance*, there is prime facie reason to think that trustworthiness is different from mere *reliability*. This is because it is possible to be merely reliable without being trustworthy. For example, my car is reliable, but I think it would be strange to claim that it is *trustworthy* in anything other than a metaphorical sense. Or consider a hit man who is committed to doing the work his mobster boss requires him to do. He is a capable killer who unfailingly completes his assignments. I think it would be appropriate to describe such a person as a reliable hit man, but I don’t think he is trustworthy. In fact part of the reason he is not trustworthy is because he is a hit man. His actions make him the kind of person we would not regard as trustworthy. And yet, certain individuals may have good reason to trust him. The mob boss who believes the hit man is committed to the mob and has goodwill toward the boss may have good reason to trust him. But persons outside the mob, especially those in rival gangs, will not have reason to do so. They may know of his reputation as a reliable killer who never misses a hit, but to them he is not trustable. They could very well be the object of his next hit. The hit man is reliable, and even trustable to some, but he is not trustworthy.

With the above constraints in hand, I argue in section 2 that trustworthiness is best understood as a kind of virtuous character. I begin by considering claims from Carolyn

McLeod (2011), Russell Hardin (1991, 1996, 2002), Annette Baier (1986, 1994), Karen Jones (K. Jones, 2011) and Philip Pettit (1995) that trustworthiness is a matter of competence and commitment to do what another is counting on you to do. On such views, persons are trustworthy when we have reason to think they are able and sufficiently motivated to do what we are counting on them to do. These views provide different explanations of trustworthy persons' motives for commitment. I show that Hardin does not require trustworthy commitment to be grounded in any specific motivations. But according to Baier, Jones and Pettit, the trustworthy are committed out of goodwill or trust-responsiveness (Baier, 1986, p. 132; K. Jones, p. 27; Pettit, 1995, p. 208). I argue that Hardin's view is too broad and fails to distinguish between trustworthiness and mere reliability but that Baier's, Jones' and Pettit's views are able to distinguish those concepts. Although as I will show, understanding trustworthiness as commitment grounded in good will or trust-responsiveness cannot account for instances where persons are competent and committed in those ways but still don't seem to be persons we would regard as trustworthy. Instead, I follow Nancy Potter in arguing that trustworthiness is best understood as indicating a kind of virtuous character. I accept her view that trustworthiness is a virtue insofar as the trustworthy care for those counting on them in ways according to a context specific mean. However, I do not accept outright the list of requirements she presents in explicating virtuous, trustworthy character. I argue that two of her requirements need to be amended. Potter requires that the institutions influencing the character of the trustworthy also be virtuous, and that the trustworthy prioritize the trust of the disempowered even when being coerced, "within certain parameters" to do otherwise (Potter, 2002, p. 29). I think these requirements are too stringent. In contrast, I hold that trustworthy persons work to resist and reform institutions that are non-virtuous, and that they are sensitive to power imbalances and seek to uphold the trust of the disempowered, but need not always be successful in doing so when coerced. As I will explain, not all trust is welcome and so the trustworthy need not be faulted for failing to uphold some trust placed in them, even when those counting on them are disempowered. But when there is a conflict between upholding the trust of the disempowered and that of dominant individuals, the trustworthy will tend to uphold the trust of the more vulnerable, disempowered individual.

In section 3 I consider the relation between trustworthiness and individuals we have good reason to trust. I identify situations where it is reasonable to trust others who either do not have virtuous character or whose character does not play a central role in another's reason for trusting them. For example, recall the Peruvian artist I discussed in *Chapter 1*. He entrusted his stall, and to some extent his physical safety, to two patrons as he delivered his artwork to their home. He may have acted foolishly, but I think it is plausible that the artist had some good reason to trust the patrons. Perhaps he thought they seemed like people of good character, but he need not have. He might have just trusted them because he thought it was in their interest to cooperate with him. A concept is needed to identify persons it is reasonable to trust regardless of their character. I use the concept of trustability to account for such persons. On my view, being trustworthy – i.e. having virtuous character – is one way, but not the only way, of being trustable – i.e. being someone whom another has good reason to trust.

I conclude section 3, and this chapter as a whole, by introducing a challenge to character-based understandings of trustworthiness. In common linguistic usage we apply the concepts of trust and trustworthiness to domains involving institutions. We may talk about trusting, or not trusting, governments, churches, universities, banks or the role-holders in those institutions. We also may talk about institutions and their role-holders as trustworthy or untrustworthy. I might say, "National Australia Bank is trustworthy" or "the Prime Minister is untrustworthy." If trustworthiness is reserved for indicating a kind of virtuous character, then we must be able to talk about the character of an institution and its role-holders. But it is not clear that institutions have character. Further, while it is perhaps clearer that role-holders have character qua individuals it is less clear how we should understand their character qua role-holders. Being a trustworthy individual may be required for being a trustworthy role-holder. But it may also be possible for roles to affect one's trustworthiness by scaffolding character. So it seems that either trustworthiness should not be understood in terms of character, or our language about institutions being trustworthy should be understood analogically. In *Chapter 3* I respond to this challenge, arguing that we can make sense of notions of trust and trustworthiness as applied to institutional contexts. I show that we can know enough about institutions and role-holders to have good reason to trust them, and I provide a positive account of institutional and role-holder character.

Section 2: From Competence and Commitment to Character

Trustworthiness involves elements of competence and commitment to do what others are counting on one to do. But those elements alone are insufficient to account for the difference between trustworthiness and mere reliability. Instead, the concept of trustworthiness should be reserved for identifying a kind of virtuous character.

Section 2.1: Trustworthiness as Competence

Competence is a basic requirement for being trustworthy. A person cannot be counted on to do something if she is incompetent to do it. To use Jones' phrase, "the incompetent deserve our trust almost as little as the malicious" (K. Jones, 1996, p. 7). Whatever else trustworthiness involves, the trustworthy are at least capable of doing what others count on them to do.

Competence can simply denote technical ability to complete a task (K. Jones, 1996, p. 7). The homeowner who entrusts her property to a house sitter must minimally take him to be capable of performing basic tasks of household management. He must at least be able to do things such as collect the post, make sure the doors are locked when he leaves, and not damage the property while the owner is away. I follow Jones in calling this, “technical competence” (K. Jones, 1996, p. 7). But competence can also involve maintaining (or adhering to) relevant social norms. For example, in addition to being able to perform basic management tasks, the house sitter will need to know the limits of his role as a temporary inhabitant of the house. If the homeowner had reason to think that he might take it upon himself to remodel portions of the property or even put the property up for sale while the owner is away, it would be reasonable for her to think him untrustworthy. In remodelling or selling the home, the house sitter violates the social norms of property ownership which extend to the homeowner but not to himself. Baier makes a similar point:

> One way in which trusted persons can fail to act as they were trusted to is by taking on the care of more than they were entrusted with – the babysitter who decides that the nursery would be improved if painted purple and sets to work to transform it, will have acted, as a babysitter, in an untrustworthy way, however great his good will. When we are trusted, we are relied upon to realize what it is for whose care we have some discretionary responsibility, and normal people can pick up the cues that indicate the limits of what is entrusted. (Baier, 1986, p. 236)

I will refer to this ability to understand and respond to social norms and expectations as “practical competence.”

The requirements of technical and practical competence vary across domains. For example, what it will mean for a *house* sitter and a *baby* sitter to be technically and practically competent will be different. While there may be some similarities in so far as both must be able to care for something another values, what that care involves is significantly different. The house sitter must only attend to basic care of another’s property while the baby sitter must temporarily care for the physical needs of an infant.

In some domains we can expect others to have an ability that is distinct from technical and practical competency. We expect them to know what kinds of *attitudes and treatment of us* are appropriate in a given situation. This will mean being practically competent in understanding what attitudes toward others are appropriate in a given situation. This is distinct from practical competency because it is not just about norm following. It is primarily focused on having an ability to know what attitudes, and expressions of them, are called for. For example, marriage partners expect each other to have an understanding of faithfulness and what it requires in a variety of situations. Or, to use an example from Jones, we expect friends to understand what it means to be loyal, kind and generous as well as what those things require across varying circumstances (K. Jones, 1996, p. 7). I follow Jones in referring to the ability to understand such requirements as "moral competency" (K. Jones, 1996, p. 7).

A satisfactory account of trustworthiness will include the elements of technical, practical and moral competency. One cannot be worthy of trust, or reliance for that matter, without at least being able to understand and do what one is being counted on to do. But competency is insufficient for trustworthiness. Consider a competent financial advisor who manipulates clients into investing in companies in which she owns shares. She has all the skills necessary to advise clients; in fact, to her clients, she seems so competent that they are easily manipulated by her. She is competent, but not trustworthy. But why such a person is not trustworthy requires explanation. One possible explanation for why the financial advisor is not trustworthy is that she is not *committed* to doing what her clients are counting on her to do. Her primary motivation is not to advise clients in the way they are counting on her to do. Or perhaps she does have the motivation to do so but has not chosen to "go with" that motivation. Instead she has let other motives win out in directing her actions—i.e. the motive to benefit financially from her clients' investments. Whatever the explanation, the financial advisor's lack of trustworthiness may come down to a lack of commitment to do what her clients are counting on her to do.

Section 2.2: Trustworthiness as Competence and Commitment

We do not only want to know that the trustworthy are capable of doing what we need them to do, but that they will in fact do it. We want to know that they are at least competent *and* committed. For the remainder of this chapter I will understand commitment broadly as aligning oneself with a given belief, attitude or course of action.

McLeod, Hardin, Baier, Jones and Pettit have each claimed that trustworthiness involves competence and commitment to do what another, or others, is counting on you to do (Baier, 1986; Hardin, 2002; K. Jones, 1996, 2011; McLeod, 2011; Pettit, 1995). McLeod presents the basic view most explicitly in saying, “Clear conditions for trustworthiness are that the trustworthy person is competent and committed to do what s/he is trusted to do” (McLeod, 2011). But as McLeod goes on to explain, not all views that understand trustworthiness as involving commitment agree on whether the trustworthy need merely to be committed or if commitment in virtue of a specific motivation is also required. Hardin holds a broad view of trustworthiness as competence and mere commitment while Baier, Jones and Pettit hold that trustworthiness requires commitment arising from specific motivations. After explaining these views, I will argue that understanding trustworthiness as competency and commitment is unsatisfactory, whether the motivations grounding commitment are specified or not. Rather, it is possible to be competent and committed – even in specific ways – without being trustworthy.

So far I have approached trustworthiness from the perspective of persons counting on others. I have been concerned with what makes us think another is trustworthy. In contrast, Hardin presents his view from the perspective of persons who want to be trustworthy and express that trustworthiness to others. According to him, persons can affect their trustworthiness by altering the interests, or “incentives”, motivating them (Hardin, 2002, p. 83). “The only way to actually affect trustworthiness is by changing one’s incentives – for example, by entering into long-term ongoing relationships with those whose trust one would like to have" (Hardin, 2002, p. 83). I take Hardin to understand interests and incentives interchangeably. This is supported by a claim of his that is similar to the one I have reproduced above but in which he talks about interests instead of incentives. "...Trustworthiness is, on the encapsulated-interest account, the result of [one] having an interest in being trustworthy toward those with whom they have ongoing interactions that are beneficial and are likely to continue to be" (Hardin, 2002, p. 84). I will follow Hardin in talking about interests and incentives interchangeably.

Hardin's view of trustworthiness is consistent with his interest-based approach to trust. As I explained in *Chapter 1*, Hardin understands trust as interaction based on the thought that another has encapsulated one's interests (Hardin, 2002, p. 1). Encapsulation of another's interests essentially changes one's interests, or incentives. For example, recall the Peruvian artist and his patrons that I presented in *Chapter 1*. It is plausible that the patrons desired to sustain their relationship with the artist, at least until he delivered their purchase. That desire could motivate them to encapsulate the artist's interests and care well for him and his stall. If that is the case, they have incentive to mind his stall well and lead him to their home without harm. Their desire to sustain interaction with the artist and encapsulate his interests changes their own interests.

In explicating trustworthiness, Hardin shows that encapsulating another's interests is not the only way to affect one's incentives. We can also change our incentives by setting up social constraints.[23]For example, pre-nuptial agreements are contractual constraints that can

[23] Onora O'Neill also thinks social constraints can be employed to establish trustworthiness. In articulating ways to encourage trust between laypersons and professionals in the medical, scientific and technological professions she writes, "Many steps can be taken to improve the trustworthiness of practices, activities and products in medicine, science and biotechnology. Fundamental ethical obligations, the rejection of coercion and deception among them, set demanding standards. Their embodiment in legislation, regulation, public policies, institutional practice and professional standards is the first and the central way of improving trustworthiness" (O'Neill, 2002a, p. 123). I will explicate O'Neill's view in *Chapter 3* when I discuss trust and trustworthiness in relationships involving institutions.

be employed to shape one's incentives. Marriage partners will most likely have reasons to remain together anyway, but pre-nuptial contracts ensure that it is in each of their interests to stay together. On Hardin's view, such contracts would make individuals trustworthy insofar as contracts can change their incentives.

> There is great trustworthiness in contracts because performance is easy to assess and enforcement is relatively easy; there is far less trustworthiness in marriage in many societies and times, because performance is too hard to measure to make enforcement work. (Hardin, 1996, p. 34)

It is contentious to claim that contracts promote trusting relationships. Contractual relationships may involve some features of trust. They can at least involve entrusting goods to another's care, being vulnerable to another and accepting that vulnerability. But the purpose of contracts is to limit the vulnerability of contractors so that they can interact, almost without having to trust each other. So while I think it is possible for some trust to occur in contractual relationships, any trust, and any of its benefits, will be constrained. As Baier writes, "Trust in fellow contractors is a limit case of trust, in which fewer risks are taken, for the sake of lesser goods" (Baier, 1986, p. 251). Even if trust can occur between contractors, it is unclear whether contracts encourage trustworthiness. For example, it is plausible that parties to a prenuptial agreement could think it unlikely that the other would violate their contract without thinking they are trustworthy. Suspicion between partners may move them to establish their contractual agreement in the first place. The presence of their contract may provide reason to think each other reliable parties to their agreement; it need not make them more trustworthy. This lack of distinction between mere reliability and trustworthiness is not limited to Hardin's consideration of contracts. The conflation of mere reliability and trustworthiness arises because Hardin does not specify the motivation grounding the commitment involved in trustworthiness.

Hardin holds that social constraints affecting incentives and so scaffolding trustworthiness are not limited to explicit contracts. Hardin gives as an example the story of

Ukifune in Lady Murasaki's *Tale of Genji* (Hardin, 1996, p. 35). Deeply troubled by a decision to choose between two partners, Ukifune opted to exit the situation all together and join a nunnery. To carry out her decision, Ukifune had to shave off her long hair, a status symbol that would take a lifetime to grow back. Hardin explains,

> Ukifune left her worldly existence by having her head shaved, taking vows, and entering a nunnery, after which she could never go back to her previous world…Ukifune therefore had little choice other than to live up to her sudden commitment to her religious vows into the distant future…In our society of radically looser conventions, you and I cannot so readily constrain ourselves as Ukifune did—our shaved heads might be nothing more than a frivolous style of the moment—and we cannot be fully believed if we assert our undying commitment to a particular religious creed or to any other purpose. We can be believed to say we are committed in this very moment—but we cannot be fully trusted to stay committed into the distant future. (Hardin, 1996, pp. 35-36)

According to Hardin, Ukifune was trustworthy because she had a whole society backing her commitment.[24] The norms of the society made Ukifune's actions so conclusive that both she and others could be certain her commitment would not be broken.

On Hardin's view then, persons can change their incentives, and so their trustworthiness, using social norms, contractual constraints or alternative mechanisms. From the perspective of persons considering the trustworthiness of others, this means looking for those who are competent and motivated to "stick with" a given course of action. And, on Hardin's view, the best reason to think another will stick with a course of action is that their incentives have been scaffolded by external constraints.

While Hardin identifies personal incentives as sufficient motivation for extended commitment, and external constraints as tools for shaping those incentives, he does not stipulate that trustworthy persons *must* be committed in any particular way. He merely points out that the most consistent commitment is grounded in incentives. His view is thus consistent with the idea that persons may be trustworthy because they are committed to a

[24] This view is evident when Hardin states, "[Ukifune] found a device that made herself trustworthy into the distant future and that simultaneously made others trustworthy not to try to change her mind" (Hardin, 1996, p. 39).

course of action out of other motives such as goodwill, or trust-responsiveness. But he thinks those motivations are less reliable than externally constrained incentives.

The problem with Hardin's view, however is that it cannot account for the difference between trustworthiness and mere reliability. It is possible to be competent and committed to doing what others are counting on you to do without being trustworthy. For example, consider a member of the U.S. House of Representatives who is committed to getting a presidential candidate elected because he knows that if she is elected, she will support initiatives he is promoting. The representative is a skilled speaker with a strong public following and so is at least technically competent to help the candidate get elected. If the candidate is aware of the representative's committed support, she will likely count on him for support in her run for office. And she would have good reason to do so because he is a reliable supporter. If trustworthiness is just competence and commitment to do what another is counting on you to do then the U.S. representative is reliable *and* trustworthy. Perhaps that is correct; maybe he is trustworthy. But on Hardin's account it is unclear what the difference would be between saying the representative is reliable and saying he is trustworthy.

Baier (1986, 1994), Jones (K. Jones) and Pettit (1995) each clarify the distinction between trustworthiness and mere reliability by qualifying the motivations for commitment. Baier characterises the commitment involved in trustworthiness as based on good will.

> The trustworthy person will feel some concern for the trusting, and this feeling will be especially noticeable if things go wrong. She will believe that she is responsible for what she is trusted for and will intend to discharge that responsibility competently and with a good grace. (Baier, 1994, p. 132)

Although Baier does not explicitly say so, I think her inclusion of the element of goodwill in trustworthiness stems from her view of trust as entrusting goods to the discretionary care of another (Baier, 1986, pp. 236-239). As I explained in C*hapter 1*, leaving things in another's discretionary care makes one especially vulnerable to “not yet noticed harm and disguised ill will” (Baier, 1986, p. 239). To have good reason to accept that vulnerability, a person will need to think a trustee will not harm her or feel ill will toward her. Rather, he will need to think she has, or will have, goodwill toward him. The presence of that goodwill can account for the difference between trustworthiness and mere reliance. Consider again the U.S. representative example. If the representative is committed to getting the presidential candidate elected out of goodwill toward her, he will not only be reliable but trustworthy to the candidate. But if, as is the case in the example, he is only committed to her cause for his own political interests, he would be merely reliable.

Baier's view is helpful in explaining the difference between trustworthiness and mere reliability, but it has been criticised as being too narrow. Jones has pointed out that we often can have reason to trust another without knowing anything about their psychological states of will: "We are often content to trust without knowing much about the psychology of the one-trusted, supposing merely that they have psychological traits sufficient to get the job done" (K. Jones, 2004, p. 4). More specifically, Baier's view encounters problems when we try to explicate the element of goodwill involved in trustworthiness. It is one thing to say trustworthy persons act out of goodwill toward those counting on them, but what that goodwill is remains unexplained in Baier's account.

In her paper "Trust as an Affective Attitude", Jones identifies two ways that Baier's notion of goodwill could be understood: goodwill as a type of friendly feeling; and goodwill as general benevolence, honesty or conscientiousness (K. Jones, 1996, p. 7). However in this paper she does not provide any further explication of the notion of goodwill. She has since analysed these two accounts of goodwill and found them wanting. First, understanding goodwill as a type of friendly feeling is, to use Jones' phrase, "far too restrictive" (K. Jones, p. 21). Jones does not explain exactly what she means by this. But I take her to be pointing to the possibility that persons could be trustworthy without having positive feelings toward those counting on them. I think it is implausible that all trustworthy persons will be positively disposed toward those counting on them. Consider a co-worker who is unfriendly but highly principled. Others in his office know they can trust him to do his part on shared projects. But they do not expect him to do so because he likes them or is particularly friendly.

Second, understanding goodwill as general benevolence, honesty or contentiousness makes the concept too broad. We are left using "goodwill" to identify some general type of positive motive which may not even be directed toward anyone in particular at all (K. Jones, p. 21). It is unlikely that this broad understanding of goodwill is what Baier has in mind, since she stipulates that the trustworthy will be motivated out of goodwill, not just toward anyone, but toward the *one counting on them*. But general benevolence, honesty and the like fails to capture that specific directedness of the attitude in question.

A third understanding of goodwill can be gleaned from Baier's explanation of the difference between trust and mere reliance. On this understanding goodwill is just characterised by motives incompatible with ill will.

> What is the difference between trusting others and merely relying on them? It seems to be reliance on their good will toward one, as distinct from their dependable habits, or only on their dependably exhibited fear, anger, or other motives compatible with ill will toward one, or on motives not directed on one at all. (Baier, 1986, p. 234)

But understanding goodwill as simply motives that are incompatible with ill will, does not bring us closer to understanding trustworthiness, for we are then left needing to explain ill will. And, as Jones points out, ill will is, at least on some readings, compatible with trustworthiness. "...On a particularly vexatious morning I find myself snarling misanthropically at the whole world, yet I can still come through for some of those who are counting on me, even if not with a smile" (K. Jones, 2011, p. 21).

Explicating exactly what goodwill means is thus not straightforward. Nevertheless, the role it plays in Baier's view of trustworthiness is still valuable. Whatever goodwill is taken to be, the purpose it serves in understanding trustworthiness is to mark a condition that provides a trustee with reason to respond to an individual counting on her. To use Jones' terms, it identifies “reason-enabling conditions” (K. Jones, 2011, p. 22). That is, conditions which provide one with reasons that influence one’s deliberation and action. Jones writes, “If I have robust goodwill towards someone, of the kind found in friendship or good collegial relations, I will take the fact that they are counting on me to be a reason in my deliberation and in my action” (K. Jones, 2011, p. 22).

On Jones’ view, taking the thought that another is counting on you as reason-enabling is characteristic of trustworthiness (K. Jones, 2011, p. 24). The trustworthy take the fact that others are counting on them to matter. This is in keeping with Jones’ 1996 account of trust as involving an expectation that another will be directly and favourably moved by the thought that you are counting on them (K. Jones, 1996, p. 6). As I explained in C*hapter 1,* Jones grounds that expectation of trust-responsiveness in an attitude of optimism about another’s goodwill and competence. Her more recent account explains why the trustworthy would be responsive to our trust: the thought that we are counting on them is *reason-enabling*. Optimism about another’s goodwill is just being optimistic that they care about the fact that we are counting on them. And that leads to an expectation of trust-responsiveness.

> There is *a* minimal sense in which the trustworthy can indeed be said to have goodwill towards the truster: just in virtue of being responsive to the fact of someone’s dependency, we *thereby* show them a measure of goodwill. The mistake is in thinking that this goodwill is something distinct from the responsiveness itself. (K. Jones, 2011, p. 24 emphasis is Jones')

I take Jones’ view of trustworthiness to involve competence and commitment grounded in responsiveness to the thought that another is counting on you. The trustworthy are able to do what we are counting on them for and committed to doing so because our counting on them is a reason-enabling condition for them.

Jones does not explain why trustworthy persons would take the fact that another is counting on them to be reason-enabling. But Pettit has offered three “mechanisms of trustworthiness” that can be employed to explicate why the trustworthy may take the fact that another is counting on them to matter (Pettit, 1995, p. 208). They are: loyalty, virtue and prudence. A person might respond to another's trust because she sees him as someone she cares about "sticking with." That is, she might respond because she is loyal to him. Someone may also be responsive to another's trust because it is in her interests to sustain her relationship with him; because it is prudent for her to do so.[25] Or, lastly, she may be responsive because of "virtue" (Pettit, 1995, p. 208). Pettit's talk about "virtue" should be kept distinct from views which understand trustworthiness as a virtue or as identifying a kind of virtuous character. As I will explain below, Potter takes such a virtue approach to trustworthiness. But Pettit does not think trustworthiness itself is a virtue or that it indicates virtuous character. Rather, his view is that trustworthy persons may be responsive to another's counting on them because of religious or moral rules and value systems they have adopted. This understanding of "virtue" is evident when Pettit writes, "Or suppose I believe that someone is virtuous: say, a god-fearing sort who can be relied upon to follow certain religious norms" (Pettit, 1995, p. 209). So as not to confuse Pettit's view with those which understand trustworthiness as a virtue itself, I will refer to the mechanism he calls "virtue" as, "being principled." The focus here is on people being trust-responsive because they think doing so would be "the right thing to do" given their context and beliefs; because of their *principles*.

Understanding the element of commitment as specifically motivated by trust-responsiveness distinguishes trustworthiness from mere reliability. And it does so without being too broad like Hardin's general commitment view, or too narrow like Baier's will-based view. Things can be merely reliable without being responsive to a truster. For example, again, my car may be reliable because it has been well engineered and because I have no

[25] In this "prudence" mechanism, I take Pettit's view to be similar to Hardin's view that persons can be strongly motivated to cooperate with others because it is in their interest to sustain a relationship with them (Hardin, 2002, pp. 83, 88).

reason to think it will break down. But my car’s reliability is not connected to its having any kind of response to the fact that I am counting on it. It may respond to my depression of the accelerator and my turning of the wheel. But its reliability is not responsive to me. It is a reliable car whether I am counting on it or not.

Non-responsive reliability can occur in persons as well. My flatmate is reliable in checking that all the doors are locked, and lights are off in our apartment when he is the last one to go to bed. But rather than checking the doors and lights in response to my counting on him, he may just do so out of habit, because of his own fear of thieves or out of a desire to save electricity.

Understanding trustworthiness as competence and commitment grounded in various types of trust-responsiveness is not without its challenges. It is possible to be competent and committed by virtue of being trust-responsive without being trustworthy. To see this, recall the example I presented in the introduction of this chapter of a hit man whose boss enlists him for an assassination. The hit man may be capable of the murder and committed to making the hit because the boss is counting on him. Further, his responsiveness to the boss may stem from one of Pettit’s three mechanisms for trustworthiness. He may be responsive because of loyalty to the boss, or out of prudence; because of what the boss might do to him should he fail. It may even be possible to understand the hit man as responsive because he is principled. He might just care about the fact that the boss is counting on him out of compliance with norms governing his mob-centred lifestyle. Given what might be called his “mob principles”, the fact that the boss is counting on him matters. In these three ways, the hit man may be competent and committed because he is trust-responsive toward the boss but that need not mean he is trustworthy. As I pointed out in the introduction of this chapter, if I (a person unconnected with the mob) know him to be a hit man, I will not trust him with very much at all and I will probably do my best to watch his every move. So while competence and commitment grounded in trust-responsiveness identifies an attribute that is distinct from mere reliability, it may not necessarily identify trustworthiness.

In addition to the challenges I have identified facing Hardin's, Baier's, Jones' and Pettit's views, understanding trustworthiness as competence and commitment misses part of what we can mean when we say someone is trustworthy. We often express something more than just that another is able and motivated in a certain way. For example if I am recommending an employee to you, I might say, "You will not need to worry about her, in my experience she has been trustworthy." In providing such a vote of confidence I am not merely stating that the employee is competent and committed to doing whatever job she is given or that she will be committed because she knows others are counting on her. I am saying something more about the *kind of person* she is. I am saying something about her character. While competence and commitment grounded in specific motivation can identify a concept that is different from mere reliability I do not think those features necessarily identify trustworthiness. In section 3 I will show that the concept which competence and commitment identify is trustability. But I develop that concept in response to a challenge facing character-based views of trustworthiness. So I turn now to first consider character-based views.

Section 2.3: Trustworthiness as a Kind of Virtuous Character

Potter provides a view of trustworthiness that is able to account for both the difference between trustworthiness and mere reliability, and for the sense in which trustworthiness marks something beyond competence and commitment. Her view is based on the premise that we are not obligated to trust but we are responsible for cultivating proper trust, which means developing an "appropriately trusting character" and cultivating "trustworthy character" (Potter, 2002, p. 12). Potter focuses on our responsibility to become trustworthy persons rather than wise trusters, but she approaches trustworthiness from the perspective of trusters.

On her view, whether we should trust or not comes down, at least part of the time, to considering another's character.

> What many of us really want to know—at least some of the time—is not just, Can I trust this person to put up flyers about the meeting next week, but: Could this person still be counted on to be trustworthy if she were backed into a corner? When faced with a conflict of loyalties, would that person still be trustworthy? When the wolf comes knocking at her door, what will she do? If the wolf comes knocking at my door, what will she do? What is involved in being trustworthy and how can we determine who is and is not worthy of our trust? (Potter, 2002, p. 13)

We want to know not just whether we have reason to expect another to behave – or even be motivated – in a certain way; we want some idea of how they will respond in changing circumstances that may test their character.

Potter uses an Aristotelian approach to explicate the importance of trustee character. She explains that in Aristotle's *Nicomachean Ethics* character is central to becoming morally good, or virtuous, persons living fully flourishing lives (Potter, 2002, pp. 13, 14). Rather than merely doing what is good by accident, the virtuous act from "enduring dispositions or states of character" (Potter, 2002, p. 14).

In addition to doing what is good, morally good – *virtuous* – persons experience appropriate feelings in conjunction with their good actions. It is these feelings, constitutive of good action, that move the virtuous toward the *mean*. That is, virtuous persons act and feel in ways that are neither excessive nor deficient in a given context (Potter, 2002, p. 15). I will refer to acting in accordance with a mean – in ways that are neither excessive nor deficient – as acting *appropriately* in a given situation or context.

Given her virtue account, Potter understands trustworthiness as a matter of having enduring character traits which dispose one to act and express feelings that are appropriate.

> *A trustworthy person, I propose, is one who can be counted on, as a matter of the sort of person he or she is, to take care of those things that others entrust to one and (following the Doctrine of the Mean) whose ways of caring are neither excessive nor deficient*." (Potter, 2002, p. 16. Emphasis is Potter's)

Potter explains what it means for the trustworthy to care in ways that are neither excessive nor deficient in terms of discretion (Potter, 2002, pp. 16-17). A person may care excessively by failing to exercise discretion about the limits of what they can reasonably care for. Or their care may be deficient if they fail to exercise discretion well regarding what sorts of things they should be entrusted with. To borrow an example from Potter, "...an excess of caring involving a lack of discretion about proper objects with which to be entrusted might be seen when one agrees to keep a confidence that ought to be reported to Child Protection agencies" (Potter, 2002, p. 17). The trustworthy can be counted on to know the limits of their ability to care and the boundaries of what they should be asked to care for. And they can be counted on to care appropriately given those parameters. On Potter's view, this does not just mean that the trustworthy exercise discretion in following a given set of norms or rules. Rather, what it means to act and feel appropriately is relationship specific. As Potter writes, "...I argue that being trustworthy is a matter of a relation between moral agents and that it doesn't quite make sense, when it comes to this virtue to talk about 'the moral agent' and her motivations as if they are independent of particular trust relations" (Potter, 2002, p. xiv). Being trustworthy then will require discretion and sensitivity to a given relational context.

On Potter's view, the kinds of persons we are, and so our trustworthiness or lack thereof, are developed in a social context. We are "situated selves" (Potter, 2002, p. 17) living with other individuals and relating to organisations and institutions. Because of this, Potter holds that trustworthiness should be analysed with reference to that social context. In *Chapter 3* I will return to questions of how best to understand trust in relations involving institutions, and the trustworthiness of institutions. For now I follow Potter in noting that social organisations and institutions help shape the kinds of persons we are and so, if trustworthiness is a matter of character, an analysis of trustworthiness should acknowledge the relevance of social environments to character (Potter, 2002, p. 17).

In *Chapter 1*, I agreed with Hardin and Baier that trust tends to be qualified to a given context (Baier, 1986, p. 236; Hardin, 2002, p. 9). We trust some persons with some things but rarely trust anyone with everything. Potter holds that a similar point can be made about trustworthiness: persons can be trustworthy to some while not being trustworthy to others. This view is evident in an example she gives of a son who is a murderer, but loves his mother very much. The mother knows about her son’s murderous actions and about his goodwill toward her.

> The son’s mother may believe her son has good will toward her and may rightly predict that he will treat her well but this certainly seems to be a delimited sense of trust—if the mother knows of her son’s murderous activities, she would be foolish to consider him trustworthy beyond the scope of their relationship (although it still may make sense to say he is trustworthy *to her*). (Potter, 2002, p. 8. Emphasis is Potters.)

The possibility of domain specific trustworthiness seems to conflict with Potter’s virtue account. Virtues are supposed to be about a whole person and so one’s trustworthiness should not be limited to certain domains. Potter acknowledges this issue: “If one can be trustworthy with respect to some people and some goods and yet not be trustworthy with respect to other people and other goods, how is trustworthiness still a virtue in the usual sense of the word?” (Potter, 2002, p. 25). To resolve this conflict Potter distinguishes what she calls “specific trustworthiness” from “full trustworthiness” (Potter, 2002, p. 25) While she continues to call persons we have reason to trust in a given situation specifically *trustworthy*, her virtue account is applied primarily to the concept of full trustworthiness. On her view, specific trustworthiness marks an attribute of being one whom another has reason to trust in a given situation; full trustworthiness identifies a kind of virtuous character.

I take Potter’s concept of specific trustworthiness to refer to the elements of competence and commitment I have identified above. This interpretation is supported by her murderous son example. Potter says that the mother has reason to think her son is trustworthy to her, though she may not have reason to think him trustworthy “beyond the scope of their relationship” (Potter, 2002, p. 8). Essentially, the son is competent and committed to his mother because of his goodwill toward her. As I have shown above, being competent and committed to doing what another is counting on you to do is not necessarily the same as being trustworthy. And so I do not think what Potter calls “specific trustworthiness” actually identifies trustworthiness. Something more is needed for one to be trustworthy. Potter’s account of ‘full trustworthiness’ is promising as it provides an additional element in identifying a kind of virtuous character; something the murderous son in her example does not have.

Potter explicates the character of the fully trustworthy, which disposes them to act in ways that contribute to human flourishing, by identifying ten “further requirements of full trustworthiness” (Potter, 2002, p. 26). These are primarily ways in which the fully trustworthy will be disposed to act and feel across varied circumstances. As I will show below, some of Potter’s requirements are too stringent. If they are accepted as they stand, very few persons, if any, will count as trustworthy. Instead, I provide more moderate versions of those requirements.

First, because trust is risky, trusters will usually place trust in another for some reason. Based on the influential approaches to trust I considered in *Chapter 1*, we might say trusters count on others because they think the risk of doing so is acceptable; because they take the other to have goodwill toward them; because they take a participant stance toward the other; or because they expect the other to be trust-responsive toward them. Each of these attitudes may make someone worthy of trust. But Potter holds that for someone to have full trustworthiness she needs not just to have encapsulated our interests, have goodwill or trust-responsiveness toward us or be a participant. Rather, she also needs to be disposed to give assurances of trustworthiness (Potter, 2002, p. 27). That is, she needs to be someone who indicates to others that they have good reason to trust her.

Second, related to providing assurances of trustworthiness is a need to take epistemic responsibility seriously (Potter, 2002, p. 27). The fully trustworthy person will be careful to consider how his own beliefs, attitudes and social context impact on his ability to relate to trusters. On this point Potter connects being virtuous with knowing well. To be fully trustworthy someone must know what another's trust requires, or might require, of him.

Third, the fully trustworthy will be sensitive to the particularities of others and understand what particular others are counting on them to do (Potter, 2002, p. 28). Potter notes that Baier has identified something like this requirement in her analysis of trust saying that trusters are relied upon to realize what another is counting on them for. Baier says that "...normal people can pick up the cues that indicate the limits of what is entrusted" (Baier, 1986, p. 236) but Potter thinks Baier oversimplifies the situation. Picking up cues can be harder than Baier suggests. It involves considering what, in the trusters' eyes, they are counting on you for. Where people are from different social backgrounds understanding what the other's trust in you involves and what it means to them can be very difficult. But the fully trustworthy will work to understand the experience of those counting on him and what exactly they are counting on him to do.

Fourth, the fully trustworthy are not just sensitive to those counting on them but also respond properly to the broken trust of others (Potter, 2002, p. 28). Part of being fully trustworthy is being aware when you have damaged the trust of another, showing that you care about that violation and working to repair the damage done.

Fifth, related to responding properly to broken trust, the fully trustworthy will deal with hurt in relationships in ways that sustain connection with others (Potter, 2002, p. 28). This trait is distinct from responding to broken trust insofar as it is meant to identify care for hurt that may not be the result of broken trust. The trustworthy will also care about hurt they have caused others as well as that which others have caused them; and work to deal with it in ways that sustain their relationships. Potter does not suggest ways that the fully trustworthy might deal with such hurt and sustain connection but I think it is reasonable to consider the following as possible ways to do so: authentically identify hurt whether one's own or another's; apologise for one's role in causing any hurt; and express to hurt individuals that you understand and sympathise with their pain.

Sixth, the institutions and governing bodies that influence individuals will need to be virtuous themselves (Potter, 2002, p. 29). As I explained above, because we are situated selves, institutions and social relations can play a role in shaping our character. Our trustworthy character, then, can be "limited or enhanced" by our institutions (Potter, 2002, p. 175). But the direction of influence is not just one way. We can also work to change institutions; making them more conducive to full trustworthiness.

> Institutional structures can promote or impede our being fully trustworthy, and so attention to betrayals of trust and responses to them, and attention to exploitation and vulnerability in terms of socially situated, particular persons, can lead to the recognition for the need to reform social institutions. (Potter, 2002, p. 29)

On Potter's view then part of *becoming* trustworthy is working to change institutions so that they do not inhibit our ability to act appropriately with regard to those counting on us. But persons are not fully trustworthy until those institutions are virtuous.

Potter acknowledges that her sixth requirement may cause cynicism about anyone attaining full trustworthiness; but she retains her view. According to her, full trustworthiness will remain "more of a vision than a reality" until institutions are reformed (Potter, 2002, p. 175). In the meantime, she understands working to reform institutions as a type of imitation of virtue (Potter, 2002, p. 176). Those resisting non-virtuous institutions will not have full trustworthiness until they succeed in their reformation but their attempts at change can express a significant move in the right direction.[26]

[26] Potter writes, "People learn to become virtuous by doing virtuous actions. Although imitating virtue is not sufficient for virtue, it is a viable start. For those in a position to effect changes in the structure of institutions

Potter's sixth requirement is too stringent. While I take her point that institutions can encourage or impede our trustworthiness, I do not think they need to be virtuous before individuals can have, and express, trustworthy character. Rather, where institutions challenge requirements for being trustworthy, resistance and attempted reformation of the inhibiting institution can express trustworthiness. Such resistance and reformation may even go further to express concern for those counting on us than if our institutions were in fact virtuous. The person who resists institutions that hinder trustworthiness shows concern for trustworthiness. She cares enough about sustaining trust to risk the consequences of deviating from powerful institutions. So while I agree with Potter that institutions and individuals can influence each other, I do not think institutions must be virtuous before persons can be trustworthy.

Potter's seventh requirement for full trustworthiness is that the fully trustworthy will "recognize the importance of being trustworthy to the disenfranchised and oppressed" (Potter, 2002, p. 29). This does not mean that the trustworthy must always uphold all trust that the disenfranchised and oppressed place in them. Rather, when upholding the trust of multiple others would *conflict*, the fully trustworthy will favour those who are already being oppressed so as not to exploit them further.[27] Potter explains that this requirement follows from the nature of trustworthiness as not being exploitative or dominating. She writes that the fully trustworthy will, "...take as a primary consideration those who are already vulnerable in

and practices, part of being trustworthy may involve a re-examination of the ways in which those institutions and practices rest on exploitation and oppression, and it may require resistance to them" (Potter, 2002, p. 176).

[27] In explaining this requirement Potter equivocates having trustworthy character with upholding the trust placed in oneself. She talks about deciding to be 'trustworthy' to someone in the sense that one decides to uphold the trust they place in oneself. She then talks about the ways in which such decisions express one's 'trustworthiness' insofar as it shows the type of person one is – that is, insofar as it shows their *character*. She writes, "Many of those conflicts take the form of having to decide to whom we most want to be trustworthy. When we can't be trustworthy to both one person and another, how should we decide whom to betray and whose trust we want to remain worthy of? Whom and what we are disposed to betray, when moral dilemmas of this sort arise, says much about our trustworthiness" (Potter, 2002, p. 29). I think the concept of trustworthiness should be reserved for marking a type of character. In section 3 I use the concept of trustability to describe persons who uphold our trust, or who we at least have good reason to think will do so, but who do not necessarily have virtuous, trustworthy character. Also, I leave Potter's mention of betrayal in the above quote to one side for now. In *Chapter 4* I show that betrayal is not just a matter of failing to uphold one's trust – as the above quote from Potter suggests. Rather, betrayal is a type of disloyalty. Specifically, I argue that betrayal is a failure to uphold normative expectations which compromises the ability of a special relationship to be sustained.

relation to dominant structures, in general, and to us, in particular" (Potter, 2002, p. 29). Further, they will do so even when they are experiencing coercion "within certain parameters" to do otherwise (Potter, 2002, p. 29).

> This claim includes, within certain parameters, situations under which we are coerced into colluding with current dominant practices to further exploit trusting disempowered others. To the extent that we allow ourselves to be coerced by dominant structures and ideologies into betraying the trust of someone who is already disenfranchised, we are failing to be trustworthy. (Potter, 2002, pp. 29- 30)

Potter does not explain her understanding of "coercion" or what she means by coercion that is "within certain parameters" (Potter, 2002, p. 29). But if the coercion she has in mind is something that we can allow to happen to ourselves, she seems to be using a weaker correlate to that concept. I think coercion is forceful; and not something that can be voluntarily accepted. Therefore, I will understand the concept Potter identifies as coercion as a weaker kind of *seduction* or *temptation* that one can allow oneself to go along with or resist. And I will understand coercion that occurs "within certain parameters" as identifying a moderate level of seduction or temptation. So, I take Potter's point to be that the trustworthy will do what they can to avoid furthering the exploitation of disempowered persons; even when dominant social structures would influence them to the contrary *to some reasonable degree*.

It can be very difficult to discern when one will be seduced or tempted by dominant systems to act in ways that further exploit the disempowered. Potter holds that to deal with these difficulties the trustworthy will develop "strategies of resistance" (Potter, 2002, p. 30). She does not explain what she takes such strategies to be. But I think they could be ways of critically assessing ideologies and structures. For example, persons may be able to develop a sensitivity to being counted on by those in dominant positions. When they are aware of such an occurrence, they may learn to question whether they are favouring the dominant over another who is disempowered and furthering the exploitation of that persons or persons. If they are in fact doing so, they could then work to resist such an exploitative outcome.

Even if ways of avoiding and resisting coercion to further exploit the disempowered are explicated, I think Potter's requirement makes trustworthiness overly difficult to attain. I agree with Potter that the trustworthy will be disposed to prioritize the trust of those who are more vulnerable when failing to do so would lead to further exploitation. However, I think it is too restrictive to require that the trustworthy always do so even when seduced or tempted to do otherwise – even when they are coerced to some reasonable degree. Rather, I propose a more moderate requirement: the trustworthy will be sensitive to the exploitation of others and seek to respond to others in ways that do not exploit, or further exploit, them. This requirement can involve working to resist coercion but it does not require the trustworthy to always overcome that coercion. The main thrust of both Potter's requirement and my moderate version of it is that the trustworthy recognize and care about the plight of those in less dominant positions and do not extend that exploitation. But my moderate requirement allows for persons to be trustworthy despite failed attempts to resist being coerced. Rather, they need only show that they care about not furthering exploitation and try to resist coercion the best they can.

Eighth, the fully trustworthy will be committed to mutuality in various relational domains (Potter, 2002, p. 80). Potter provides two reasons for including this trait in her understanding of full trustworthiness. First, mutuality is needed to foster and sustain interaction between persons: "We need just, responsive, democratic interaction in our interpersonal relationships as much as we do in civic life, and in neither domain should they be assumed to be secured or circumscribed" (Potter, 2002, p. 30). In order to continue relating together persons must be able to confidently call into question imbalances in their relationships. The second reason Potter includes mutuality in full trustworthiness is that persons need to cooperate in order to live flourishing lives, and one form of cooperation is trust. She holds that non-mutuality hinders our ability to trust each other and cooperate, which inhibits flourishing (Potter, 2002, p. 30). The fully trustworthy person then, will work to encourage mutuality in relations for the purpose of sustaining interaction and fostering human flourishing.

Ninth, the fully trustworthy, "*work to sustain connection in intimate relationships while neither privatizing nor endangering mutual flourishing*" (Potter, 2002, p. 30 Emphasis is Potter's). Relations between intimates are characterised by openness and closeness. Such transparency and connection can enhance one's life enabling her to flourish more *within* intimate relations but also *beyond* them. As Potter writes, "Being trustworthy in intimacy in the ways that sustain connection allows us to expand our capacity for caring, just, and mutually enlivening relationships beyond primary ones to our social, professional, and political lives in civil society" (Potter, 2002, p. 31).

Tenth, the fully trustworthy will also need to have other virtues (Potter, 2002, p. 31). Potter takes trustworthiness to be part of a "family of virtues" that are other regarding and altruistic (Potter, 2002, p. 31). In this "family" are trustworthiness and, among other virtues, compassion, thoughtfulness, beneficence and justice. Potter notes that it may seem strange to think that someone who fails to express one of these other virtues would also fail to be trustworthy: "Certainly, one may argue, Alan can be trustworthy to Bill with regard to some x (for example, where x = posting Bill's letter, or returning Bills' borrowed car) without considerations about Alan's disposition toward compassion entering in" (Potter, 2002, p. 31). But Potter explains that on an Aristotelian account like hers, such dispositions matter to the expression of other virtues such as trustworthiness. This is because trustworthiness involves acting and feeling in accordance with a context specific mean. What counts as appropriate action and feelings in a given context will not just involve undertaking what another is counting on you for (i.e. posting Bill's letter or returning his car) but doing so in the right way, given a certain situation. So, in Potter's example, if Alan posts Bill's letter or returns his car at the expense of being compassionate to someone else – Potter uses the example of Alan ignoring an elderly man in need just so he can post the letter on time – he fails to express trustworthiness (Potter, 2002, pp. 31-32). Alan still does what Bill is counting on him to do but does not show himself to be fully trustworthy.

On Potter's view then, full trustworthiness is a matter of being the kind of person – having the kind of character – who recognizes and cares about the fact that others are counting on you and who responds appropriately to that dependence given a specific context. But that is not all; it is having the kind of character such that the trait of being appropriately responsive to the trust of others is supported by the rest of one's virtuous character. Understanding trustworthiness in Potter's "full" sense is able to account for the difference between trustworthiness and mere reliability. The trustworthy are not merely competent and committed to doing what we count on them to do but, rather, have virtuous character disposing them to act in ways that contribute to human flourishing. Potter's view can also account for the sense in which identifying someone as trustworthy says something about them as a whole person. For these reasons, I think that Potter's character-based view provides an important account of trustworthiness. But as I have explained above, some of her requirements for full trustworthiness need to be amended. With those changes made, trustworthiness is still a virtue, but what it means to have character expressing that virtue is slightly different. The trustworthy are the kinds of people who give assurances of their trustworthiness; take epistemic responsibility seriously; are sensitive to the particularities of others; respond properly to broken trust; deal with hurt in relations in ways that sustain connection with others; work to resist and reform non-virtuous institutions; are sensitive to the exploitation of others and seek to respond to trusters in ways that do not further any exploitation; are committed to mutuality; work to sustain connection in intimate relationships without endangering mutual flourishing; and have a number of "other regarding" virtues.

Understanding trustworthiness as a virtue has not gone unchallenged. Jones has cautioned against taking such an approach for four reasons: persons need not be at fault for, “...refusing to respond to unsolicited trust with trustworthiness...”[28]; trustworthiness can be employed to meet good and bad ends; when trust is in the service of bad ends, “trust busting”, should be encouraged or required[29]; and it can be right to elicit trust and then “bust” it with treachery (K. Jones, 2011, pp. 38-39). But as I will now explain, Jones’ concerns miss their mark because she assumes an understanding of trustworthiness that is closer to Potter’s concept of specific trustworthiness than to full trustworthiness.

Regarding Jones' first concern, I agree that persons are not always at fault for refusing to commit to carrying out what they are being trusted to do. As Jones says, “Sometimes, by their trust, others can attempt to manipulate you into responding to their dependency, and it need be no fault on your part if you refuse to succumb to their pressure” (K. Jones, 2011, p. 38). But on Potter’s virtue account, trustworthiness is not merely carrying out what others trust one to do. To put it simply, instead trustworthiness is doing what is appropriate in a given situation. And that may mean refusing to commit to doing what another is counting on you to do or committing but then foregoing that commitment. For example, recall Potter’s example of Alan who is trusted to post a letter for Bill. If Alan stops and acts out of compassion for an elderly man rather than continuing on to post Bill’s letter, he forfeits doing what Bill trusted him to do. But he still expresses trustworthiness.

It may seem that Potter’s seventh requirement for full trustworthiness conflicts with Jones’ concern. That requirement is that the trustworthy will prioritize upholding the trust of those in positions of heightened vulnerability (Potter, 2002, p. 29). But as I pointed out when explaining that requirement, the trait Potter is identifying is not a disposition to uphold *all*

[28] Jones explains this point more fully in “Trust as an Affective Attitude.” There she explains that sometimes we may rightly feel that we cannot live up to another’s trusting expectations of us and so we may refuse to respond with trustworthiness (K. Jones, 1996, p. 9).

[29] Baier first made this point (Baier, 1986, p. 232).

trust placed in oneself. It is being disposed to prioritize the trust of the especially vulnerable when upholding their trust conflicts with carrying out what others in more powerful positions are counting on you to do. I do not see any reason for Potter's seventh requirement, or my moderate version of it, to fault the trustworthy for refusing to carry out what another is trusting them to do; whether that other is disenfranchised or not.

If trustworthiness is understood as a singular trait, like trust-responsiveness for instance, I think Jones's second concern would also be correct. If trustworthiness is just being responsive to the thought that another is counting on you, it is possible for trustworthiness to be employed to meet bad ends. If a friend counts on me to care for his child while he is at work, my trust-responsiveness to him will most likely be employed for good ends. But if a drug dealer counts on me to help her smuggle contraband onto a commercial flight, my trust-responsiveness to her can be used to meet bad ends. Jones' criticism assumes that whatever trustworthiness is, it is value neutral and so is not only expressed in actions meeting good ends but also those meeting bad ends. But again, on Potter's view full virtuous trustworthiness is not just doing what others are counting on you to do but responding to their trust appropriately in ways that encourage flourishing. If I give in to the drug dealer and commit to helping her smuggle contraband, on Potter's account I would not be expressing trustworthiness. I would be competent and committed to helping the drug dealer reach bad ends, but I would not be showing that I could be trusted because of the kind of person I am. I would not be expressing concern for what is good and contributes to human flourishing. In fact, full trustworthiness would probably call for me to bust the dealer's trust in me.[30]

On her third and fourth points, Jones seems to understand trust-busting as contrary to

[30] Amy Mullin has also argued against trustworthiness being a virtue, claiming that it involves strength of character but that such strength can be, as she says, "put in the service of immoral action" (Mullin, 2005, p. 322). But like Jones' concern, Mullin's critique misses the point of Potter's view, since on that view trustworthiness is not just strength of character but a consistent kind of character. And on Potter's view, trustworthy persons would presumably have the kind of consistent, virtuous character that would resist immoral action.

trustworthiness so that instances where it is good to bust trust are counter-examples to the claim that trustworthiness is always good – and so virtuous. But as the drug dealer example above shows, it is not clear that trust-busting is contrary to trustworthiness. Another example of someone who expresses trustworthiness while busting trust is an ethical whistle blower. The person(s) who the whistle-blower has identified as taking part in corrupt actions may have reason not to trust the whistle-blower with information about corrupt dealings in the future, but it seems to me that such trust would be corrupt itself. So while blowing the whistle may bust trust with those individuals, it can strengthen others' trust in the whistle-blower as someone who can be counted on to not comply with corrupt dealings. As such, trust-busting can express trustworthiness. And this is true whether the trust is merely busted, as in the drug dealer case and the whistle-blower case, or if it is actively elicited and then busted. Consider an alternate drug dealer example of an undercover police officer who works her way into the centre of a drug ring. The officer encourages the criminals to trust her only to later break that trust for reasons of law enforcement. In such a case it is not wrong to actively elicit trust and then bust it; and doing so need not conflict with Potter's account of full trustworthiness as indicating virtuous character. In eliciting and busting the drug dealer's trust, the undercover officer works to establish a good society where the presence of deception and harmful drugs is decreased.

When trustworthiness is understood as a kind of character involving integrity with other virtues it is not susceptible to Jones' concerns. As I have tried to show, Jones points to some correct facts: there need be no fault in failing to fulfil another's trust in you; commitment to what another is counting on you for can be employed to reach bad ends; we can be required to bust trust; and it can even be good to elicit trust and then bust it with treachery. Those considerations, however, need not mean that trustworthiness ceases being a virtue. Trustworthiness needs to be distinguished from merely being competent and committed to doing what another trusts you for. Potter's account of full trustworthiness does just that and so her character-based view, with my amendments, provides a satisfactory understanding of trustworthiness. It is able to account for the difference between mere reliability and trustworthiness, and for the sense in which trustworthiness identifies something beyond competence and commitment. Also, it does not make trustworthiness unattainable. And it is not susceptible to Jones' concerns.

To recapitulate, there is reason to think trustworthy persons are at least competent and committed to doing that which others are counting on them to do. But such competence and commitment may indicate mere reliability rather than trustworthiness. Positing a specific type of motivation for commitment does help to distinguish trustworthiness from mere competence and commitment, but even when a satisfactory motivation is found the sense in which trustworthiness marks something about a whole person is unaccounted for. Understanding trustworthiness in terms of virtuous character enables us to explain both the difference between trustworthiness and mere reliability and the sense in which to regard someone as trustworthy is to identify something about the whole person.

With my amended version of Potter's character-based account in hand, I now turn to consider whether trust is only reasonably placed when its object is a trustworthy person. I will argue that trustworthy persons are not the only ones it is reasonable to trust, and will introduce the concept of trustability as a broad category that identifies those persons that we do have good reason to trust. Trustworthiness, I suggest, is one type of trustability.

Section 3: Supplementing Trustworthiness with Trustability

While there is good reason to reserve the concept of trustworthiness for indicating a kind of character, not all trust involves counting on persons who are thought to be working toward full trustworthiness. For example, developmental trust involves trusting another despite thinking their character is deficient. Recall the example of the parent who entrusts the family car to her teenage son. The parent does not trust the teenager with the car because she thinks he has good character. Rather, at least part of the reason she trusts the teenager is to foster the development of his character. But while developmental trust is not based on a trustee having virtuous character, it can still be done for good reason. The parent's hopes of developing the teenager's character can provide good reason for her to trust him. And she can have good reason to do so if she expects the teen will be responsive to her counting on him. But as I argued in section 2.2, this expectation of trust-responsiveness is not the same as an ascription of trustworthiness. So it is possible to have good reason to trust the teen despite thinking he does not have the kind of character required for trustworthiness.

Trusters who know little about another's character can also have good reason to trust. This will be the case in instances of reasonable trust placed in strangers. Consider again the example of the Peruvian artist that I presented in C*hapter 1*. Having only just met the patrons, the artist would not have known much about their characters; they were after all mere strangers to him. And yet, it is unlikely that all trust between strangers is unreasonable. Rather, as I explained in *Chapter 1*, persons may trust strangers when they believe the strangers have goodwill toward them or because they think it is in the stranger's interest to continue interacting with them. The artist might have thought the patrons seemed to have nothing against him and so took himself to have good reason to trust them. Or he might have thought that it was in the patrons' interest to do whatever was required to close the art deal and so trusted them. As I showed in section 2.2, consideration of goodwill and interests is insufficient for explaining trustworthiness. But my point here is that those considerations can still provide one with good reason to trust another.

Trust in collectives can also be reasonable despite not being based on an assessment of trustee character. This is because it is highly unlikely for members of a collective to have similar character in general or for them all to have virtuous, trustworthy character.[31] And yet, persons can have good reason to trust collectives. In the "trust fall" example I introduced in *Chapter 1*, the blind-folded participant can have good reason to trust the others to catch her despite them having different kinds of character. The blind-folded participant may just trust because she thinks the others will be responsive to the thought that she is counting on them to catch her. Competence and commitment grounded in trust-responsiveness can still provide persons, like the blind-folded "trust fall" participant, with good reason for trusting others.

As the above examples show, persons with virtuous character do not exhaust the class of those we can have good reason to trust. If the concept of trustworthiness is used to identify those it is reasonable to trust, this result would count against the character-based understanding for which I have argued. It would seem that it is not just persons with virtuous character who should be regarded as trustworthy. One solution to this challenge would be to accept Potter's talk of "full" and "specific" trustworthiness at face value – as identifying two different kinds of trustworthiness. Then reasonable trust could be that which is placed in specifically or fully trustworthy persons. But as I showed in section 2.3 there are good reasons for not taking this approach. Specific trustworthiness seems limited to identifying traits of competence and commitment and that is not all we mean when we identify someone as trustworthy. Instead I think reasonable trust should *not* be understood as just that which is placed in trustworthy individuals. While persons do have good reason to trust those with virtuous character, trustworthiness so understood is not a necessary condition for reasonable trust. Jones is right when she says that, "Trust is a fitting response to trustworthiness and the trustworthy are fit objects for our trust" (K. Jones, 2011, p. 2). But the trustworthy are not the

[31] This is not to say that organizations involving groups of people cannot be trustworthy. In *Chapter 3* I explain how institutions can be trustworthy. But I take what Todd Jones has termed a "non-summative" approach where organizations are not understood as just the sums of their members (T. Jones, 2007, p. 446). I argue that trustworthy institutional character is not just the sum of the trustworthiness of its members. So my claim here that it is unlikely that members of groups will not all have trustworthy character need not conflict with my claim in *Chapter 3* that the concept of trustworthiness is applicable to the institutional context.

only fit objects of our trust.

Since it can be reasonable to trust someone who is not necessarily trustworthy, what should we call those we have good reason to trust but who do not have virtuous character? If they are not just trustworthy, perhaps they are merely reliable. But that is incorrect because it is possible for someone to be merely reliable without being a person that others have good reason to trust. Consider again Potter's example of the murderous son who loves his mother. If he can be counted on to do whatever he sets his mind to, we might say he is reliable. But killing is what he sets his mind to which counts against him having virtuous character. He is reliable but not trustworthy. And yet his mother may have good reason to trust him. After all, she knows without any doubt that her son has goodwill toward her.

To capture the attribute of being a person whom another has good reason to trust I introduce the concept of *trustability.* Unlike trustworthiness, I do not take trustability to be a matter of character. The trustworthy are also trustable but trustability is more about the way one is seen by a truster. For example, the murderous son is trustable to his mother but probably not to the rest of us. That is because she knows him to have goodwill toward her and so sees him as someone that it is reasonable for her to trust. But I do not see the son as someone that I have good reason to trust. If I thought he had goodwill toward me I might change my mind and take him to be trustable. Still, his having goodwill towards me need not be the only thing that could make me think he is trustable. Persons may be trustable for various reasons. They may have goodwill toward those counting on them, but they need not. We may just expect them to be trust-responsive for some other reason. Here Pettit's mechanisms for "trustworthiness" can be seen to be devices for trustability. I may have reason to accept vulnerability and trust another because I think he will be loyal to me. I may do so because I think he follows a moral or religious code which would not allow him to violate my trust. Or I may think it is prudent for him to respond well to my counting on him and so have reason to place trust.

In addition to trustee will and responsiveness, trustees can be trustable because of their interests. In section 2.2, I discussed Hardin's view that trustee incentives can scaffold trustworthiness. I said that while Hardin specified incentives as motivations for trustee commitment his view is consistent with understanding trustworthiness as competence and commitment grounded in unspecified motives. On his view all that really matters is that a trustee will keep their commitment for some reason. I argued that such an understanding of trustworthiness is unable to account for the difference between mere reliability and trustworthiness. But that need not mean that one's incentives cannot provide others with reason to see one as trustable. Consider again the Peruvian artist example from C*hapter 1*. It would be understandable for the artist to think that the patrons had an incentive not to harm him or his stall. After all, they needed him to get the artwork safely to their home. On the analysis I have provided in this chapter, this incentive does not make the patrons trustworthy. But it can provide the artist with reason to accept vulnerability to the patrons and trust them. In short, one may be trustable because of any reason others have for thinking she will not violate their vulnerability – i.e. because she is competent and committed (because of various motives) or has trustworthy, virtuous character.

Supplementing the concept of trustworthiness with the concept of trustability explains some of the limitations of the accounts discussed above, which characterise trustworthiness in terms of competence and specifically motivated commitment. As I showed, competence and commitment, even when motivated by trust-responsiveness, is not just mere reliability but is also not trustworthiness. That is, another's being competent and committed because of trust-responsiveness makes them more than merely reliable but it does not necessarily make them trustworthy. Instead, it makes them trustable. And so I have explained that we can in fact have good reason to trust persons even if they are not trustworthy.

I began this chapter by saying that because trust is risky we must have a sense of who it is reasonable to trust. I assumed that persons could at least have good reason to trust the trustworthy but that the concept of trustworthiness needed to be explained before that answer would be informative. I have argued for an understanding of trustworthiness as a kind of virtuous character. But I have also argued that the trustworthy are not the only ones we can have good reason to trust. Rather, it is reasonable to trust those we think will not violate our vulnerability to them. And there are a number of reasons we may think others will not do so including, but not limited to, believing they have goodwill toward us; expecting them to be trust-responsive; or thinking it is in their interest to uphold our trust.

In this chapter I have limited my discussion of trustworthiness mostly to trust relations between individuals. But in addition to trusting individuals, we also place trust in groups of individuals and in institutions. For example, I may trust a committee to construct a satisfactory code of ethics for the philosophy department. Or I might trust the police department to keep the streets of my neighbourhood relatively crime-free at night. Sometimes we even talk about such complex trustees in terms of trustworthiness. We might say, “Royal North Shore Hospital is trustworthy.” Or, “The Australian Department of Immigration is not trustworthy.” But it is unclear whether such language should be taken at face value or if it is just metaphor. Groups and institutions may not be the kinds of things that can have character and so, on the view of trustworthiness I have argued for in this chapter, could not actually be trustworthy. So whether the character-based approach to trustworthiness I have argued for is applicable to trust between individuals and institutions needs to be explained. Further, if they cannot be trustworthy, groups and institutions may, in theory at least, be trustable to us. But it may turn out that we cannot know enough about groups or institutions to have any reason to trust them.

In *Chapter 3* I analyse relations between individuals and groups and institutions and argue that we can give substantive accounts of trust and trustworthiness in such domains. I show that interactions involving groups or institutions can be sufficiently similar to paradigm cases of trust and so should be understood as instances of trust. Further, I show that it is possible to know enough about groups, institutions and the role-holders in them to see them as trustable. In addition, I follow Todd Jones (2007) in providing a functional account of institutional character and show how such character can be virtuous. This provides me with a substantive account of trust and trustworthiness at the institutional level that can be used to develop a broad analysis of betrayal in C*hapter 4*.

Chapter 3: Institutional Trust and Trustworthiness

Section 1: Introduction

My analysis of trust and trustworthiness in C*hapters 1* and *2* focused on interpersonal relations between individuals. But persons can be at risk of having trust violated not only by other individuals but also by institutions[32] and their role-holders. This is evident in cases of abuse at the hands of Catholic priests. In *Broken Trust: Stories of Pain, Hope, and Healing from Clerical Abuse Survivors and Abusers* (Fleming, 2007), priests and parishioners provide firsthand accounts of their ordeals of abuse. Consider the testimony of one survivor who uses trust-relevant terms in writing about a life of extended abuse.

> This story begins with the loss of a sibling, but moves very slowly and subtly into one of emotional and sexual abuse. I was dominated and manipulated and felt shame, guilt and fear. I was the victim of misuse of pastoral authority, the breach of a sacred *trust*. But this is also the story of love, grief, healing, friendship, and forgiveness. A priest abused me and put me into servitude for fourteen years. This is a story of my journey toward wholeness. My willingness to share this story is driven by a hope that other women who have been victimised by the overwhelming abuse of power within the Catholic Church might finally have a voice, a voice that is long overdue for women in the Church, for I believe our numbers are many (Fleming, 2007, p. 141. Emphasis added.).

This survivor's choice of words is significant. She speaks of abuse at the hands of a role-holder – the priest – as an abuse of *trust*. She trusted the priest qua role-holder and had that trust severely broken. But her trust is not only broken by a role-holder. The survivor's story also involves violation by an *institution* within which the interaction between her and the priest was situated. That she understands her own story in this way is evident in a list she

[32] I understand institutions broadly as social organizations. As such, institutions can vary in size. Small businesses and neighbourhood clubs may count as social organizations and large businesses, universities, governments and religious groups will as well. Institutions are also internally differentiated. Persons may trust some parts of institutions but not others or only some of an institution's role-holders. For example, I may count on the Australian Government or I may count on the Department of Immigration and Citizenship (DIAC) within the Australian Government. Throughout this chapter I will speak about 'trust in institutions' and, when relevant, specify what types of institutions, and which parts of them, are the focus of my analysis.

writes expressing her anger. “I’m angry because I placed so much trust in him and that trust was betrayed [...] I’m angry at the Church for not taking responsibility for its clergy [...] I am angry at the institutional Church for being so incredibly dysfunctional” (Fleming, 2007, p. 153).

The above example shows the relevance and importance of trust, trustability and trustworthiness in institutional contexts. I do not think the survivor’s story would be accurately expressed if these concepts were not mentioned in the explanation of what happened to her. She presumed the Catholic Church and its role-holders to be trustable, if not trustworthy; she trusted and had that “sacred” trust violated (Fleming, 2007, p. 141). And neither the Catholic Church nor the priest responded well to that broken trust.

There are three challenges to applying concepts of trust, trustability and trustworthiness to institutional contexts. First, in C*hapter 1* I argued that trust is best understood as a cluster concept. I explained that to count as trust, specific instances of interaction must be sufficiently similar to paradigm cases involving trust’s central features. Although relations among individuals, institutions and role-holders are similar in some respects to interpersonal relations, they may be insufficiently similar to paradigm cases of trust to be instances of trust themselves.

Second, Russell Hardin has argued that a significant difference between interpersonal relations and those in the institutional context is that persons cannot know enough about institutions and role-holders to place reasonable trust in them. Instead, he holds that seeming instances of trust involving institutions and role-holders should be understood as “quasi-trust” (Hardin, 2002, p. 157); a type of dependence based on inductive prediction. Further, his view suggests that institutions and role-holders cannot be trustable. As I explained the concept of trustability in C*hapter 2,* to be trustable is to be someone whom another has good reason to trust. On Hardin’s view, we cannot have good reason to trust institutions or their role-holders because we cannot have the requisite knowledge about them.

Third, at first glance, the character-based view of trustworthiness that I argued for in *Chapter 2* seems inapplicable to institutions and role-holders. On my view, trustworthiness is characterised by a kind of integrated, virtuous character. But it is unclear that institutions have character at all. Further, if institutions can have character, what it means for institutions and role-holders to have integrated, virtuous character needs to be explained.

In this chapter, I respond to these challenges and show that it is possible to provide substantive accounts of institutional trust, trustability and trustworthiness. In section 2 I show that the cluster concept of trust is, at least in theory, applicable to the institutional context: relations involving institutions and role-holders can be sufficiently similar to paradigm cases of trust to count as instances of trust themselves. But while the concept of trust is in theory applicable to the institutional context, it may turn out that, practically, persons cannot know enough about institutions and their role-holders ever to have good reason to trust them.

In section 3 I show that persons *can* know enough about institutions and role-holders to have good reason to trust them. As I explain, persons can at least have knowledge relevant to trust in the institutional context. They can know enough to distrust institutions and their role-holders. But that is not all; persons can also know enough to have good reason to trust institutions and their role-holders. That is, institutions and role-holders can be trustable. In institutional contexts trustability requires that individuals have reason to think institutions and/or role-holders are competent and committed to doing what they are counting on them to do. I consider Onora O’Neill’s views on accountability and transparency as mechanisms that encourage trust. I argue that the condition she refers to as “intelligent accountability” (O'Neill, 2002b, p. 58) can encourage institutions and their role-holders to be trustable while what she calls “active checking” (O'Neill, 2002b, p. 77) can aid individuals in knowing whether institutions and role-holders really are competent and committed to doing what they are counting on them to do. If an institution and/or role-holder is made accountable to the right people and for the right things, and if trusters are in a position to check the information they receive about an institution and/or role-holder, then trusters can also be in a position to know whether or not the institution is competent and committed to doing what they are counting on them to do.

As I explained in *Chapter 2*, it is possible to be trustable without being trustworthy. In section 4 of this chapter I argue that institutions and role-holders can be not only trustable but also trustworthy and that the character-based account of trustworthiness I argued for in *Chapter 2* can be applied to institutional contexts. In section 4.1 I show that once an institution is understood functionally, traits can be attributed to it based on the way it fulfils its functions. Those traits, along with associated institutional roles and policies can then be used to identify an institution's character. In section 4.2 I present Justin Oakley and Dean Cocking's view that virtuous professional roles are those aiding the fulfilment of professional functions which contribute to human flourishing. In section 4.3 I use Oakley and Cocking's view and the requirements for trustworthiness I adopted from Nancy Potter in *Chapter 2* to develop an account of institutional trustworthy character. I argue that trustworthy institutions are characteristically disposed to contribute to human flourishing: the institution's roles, role-holders and policies dispose them to fulfil functions which, when performed well, contribute to human flourishing. I use Potter's requirements to explicate what it means for institutions to be disposed to perform functions well. By having policies for responding to broken trust; responding to others in their particularity; and, among others, responding to the disempowered and exploited in a way that does not further their disempowerment or exploitation, trustworthy institutions are disposed to conduct themselves well while performing functions which contribute to human flourishing.

Section 2: Extending the Concept of Trust to Institutional Contexts

The cluster concept of trust, I claim, is applicable to relations involving institutions and their role-holders. To show this, I argue against what I will call “Hardin’s conceptual limitations claim”: that the concept of trust does not generalize from relations between individuals to those involving institutions (Hardin, 2002, pp. 151-157). As I will explain, this claim is based on what I will refer to as “Hardin’s epistemic limitations claim”: that individuals cannot know enough about institutions or role-holders to trust them. After explicating these challenges, I will show that relations among institutions, role-holders and individuals (external and internal to institutions) with whom they interact can be sufficiently similar to paradigm cases of trust to count as instances of trust themselves.

Section 2.1: Challenges to Applying the Concept of Trust to Institutional Contexts

Hardin’s analysis of trust in institutional contexts is situated in a discussion of whether governments need the trust of their citizens in order to function well (Hardin, 2002, p. 151). Because he does not think the concept of trust generalizes from interpersonal relations to those involving institutions, he reasons that government institutions must not need the trust of the people in order to function well.

While Hardin’s view is motivated by examples involving governments, it raises broader challenges about whether the concept of trust can be applied at all to relations involving institutions and role-holders. I set aside the question of whether governments need their citizens’ trust in order to function well. What I am primarily concerned with here is whether the concept of trust is applicable to institutional contexts at all, that is, whether Hardin’s conceptual limitations claim is correct.

As I explained in C*hapter 1*, Hardin takes an interest-based risk-assessment approach to trust. On his view, trust involves interacting with another because you think the other has encapsulated, or will encapsulate your interests into their own. Hardin thinks that trust does not generalize to the institutional context in part because *interest* does not generalize from individuals to institutions:

> If our notion of trust comes from understandings of individual behaviour and character, the term is likely to be entirely out of place in application to a nation, group, or institution. There may be ways to interpret the notion to apply it to such actors, but it is not likely to be prima facie applicable without interpretation. [...] It is now a commonplace understanding that interest is not readily generalized from individual to group or national levels. It should not surprise us to find that trust, which is commonly at issue just because interests are at stake, is not readily generalizable, either. (Hardin, 2002, p. 153)[33]

Hardin does not just think trust fails to generalize to institutional contexts because interests do not generalize from individuals to institutions. Rather, even if interests did generalize to institutions, he does not think we could have the knowledge necessary for trusting them. I now turn to explicate this epistemic limitations claim.

On Hardin's view of trust, the primary reason for thinking another is likely to encapsulate your interests is because they are motivated to sustain their relationship with you (Hardin, 2002, pp. 3, 13). But Hardin holds that we have no reason to think an institution, or its role-holders, have encapsulated, or will encapsulate, our specific interests (Hardin, 2002, p. 153). This is primarily because he thinks we cannot have sufficient knowledge of their motives. "In the encapsulated-interest account, I must know that the agents or the institution

[33] I think it is possible for some interests to generalize from individuals to institutions. For example it seems there will be some convergence between the interests of a national government and the citizens whose general interests they are supposed to care for. Still, it is not clear that specific interests of particular citizens will generalize to the national institutions meant to care for them. And so I will accept Hardin's claim that interests do not generalize to institutions.

Even if interests did generalize to groups and institutions, Hardin's interest-based view of trust would be problematic. In *Chapter 1* I showed that Hardin's view was unable to account for the difference between trust and mere reliance. I showed this by considering the example of the female employee who relies on her sexist boss for job security but does not trust him. Consider an alternate version of that sexist employer example. Instead of relying specifically on a single boss for job security, the female employee may rely on an institution for job security. Perhaps she is employed as an academic philosopher by a university that has a male-dominated culture of sexism against women. In her experience, the university's administrative staff tends to discount proposals for reform suggested by women and treats female staff as second-rate academics. She suspects that the university only sustains the employment of the few women in its ranks in order to meet minimal equal opportunity standards and not be shunned by others as a chauvinistic academy. It is in the university's interests to sustain the employment of its female academics, but it does not really care about diversity. In this environment, the female philosopher may have reason to rely on the university for job security but will most likely not trust it. So even if Hardin thought his interest-based account did apply to the institutional context, it would fail to distinguish between mere reliance and trust.

act on my behalf because they wish to maintain their relationships with me. That is generally not possible for government and its officials" (Hardin, 2002, p. 156). I think Hardin's claim that persons cannot know whether governments or officials are motivated to maintain relationships with them is incorrect. Rather, some people *can* be reasonably certain that government officials are motivated to sustain their relationships with them. For example, Christian leaders in America might have good reason to think that government officials running for election are motivated to sustain relationships with them in order to gain their support and influence in winning "the Christian vote." Persons can also have reason to think institutions and role-holders are motivated to sustain relations with them outside of the governmental context. For example, a given store and its salespersons may be motivated to sustain customer-relations with customers. If I register a complaint about a faulty product with a company's customer-relations department, it is plausible that the department personnel will respond to my complaint because they are motivated to maintain good relations with their customers.

Even if we had no reason to think that an institution or its role-holders were motivated to sustain their relationship with us, we can still have sufficient knowledge about institutions and role-holders to have good reason to trust them. Institutions and role-holders can in fact be trustable to us. After explaining how the concept of trust can be applied to the institutional context, in section 3 I will return to explicate institutional and role-holder trustability.

Hardin considers taking an approach to trust in institutional contexts that is less demanding than his encapsulated-interest account. I will refer to this as a *summative approach* – where trust in an institution correlates to the *sum* of one's confidence in that institution's members or role-holders (Hardin, 2002, p. 156).[34] On this approach, trust in an

[34] My use of the term "summative" comes from Reza Lahroodi's' work on collective epistemology. In "Collective Epistemic Virtues" Lahroodi writes, "*Simple summativism* asserts that a group *G* has the trait *T* if, and only if, all or most members of *G* have *T*" (Lahroodi, 2007, p. 285. Emphasis is Lahroodi's). I discuss Lahroodi's work further in section 4 when explicating institutional character.

institution will just be trust in its role-holders. For example, when we talk about trusting a branch of the Australian Government, it could be thought that what we really mean is that we trust the individuals that make up that branch.

Hardin doubts that trusters can have sufficient knowledge about enough institutional role-holders to be confident about their interests and to trust them. As he writes,

> Virtually no one can know enough of the large number of individual role holders to claim to be confident of judging that these role holders have interests or the relevant moral commitments to do what would serve their client's interests. (Hardin, 2002, p. 156)

Because Hardin thinks our knowledge about collectives of role-holders is insufficient, he thinks our knowledge about institutions remains insufficient. I will explain in section 4 that there are other problems with taking a summative approach as well – primarily that institutions are not always reducible to the sum of their parts, i.e. their role-holders.

Rather than taking a summative approach, Hardin suggests that talk of trust placed in institutions should be understood as what he calls "quasi trust" – a reduction of trust to confident expectation (Hardin, 2002, pp. 156-159). As he explains it, quasi trust is dependence "grounded in inductive extrapolation from past behaviour or reputation" (Hardin, 2002, p. 157). That is, on Hardin's view, trust in institutions amounts to expectations of what an institution, or its role-holders, will do based on what they have done in the past. Hardin is not saying that we do not count on institutions or that we should always adopt attitudes of scepticism toward institutions, but he is claiming that we should call our attitudes toward institutions and their role-holders "confidence" and, at most, "quasi-trust" rather than trust (Hardin, 2002, pp. 156-157).

I disagree with Hardin. I think it really is trust that can be at stake in relations involving individuals, institutions and role-holders. Consider the example of the Catholic Church with which I began this chapter. Recall how the survivor of abuse expressed her anger toward the priest who abused her and toward the church: "I'm angry because I placed so much trust in him and that trust was betrayed [...] I'm angry at the Church for not taking responsibility for its clergy" (Fleming, 2007, p. 153) and "I am angry at the institutional Church for being so incredibly dysfunctional" (Fleming, 2007, p. 153). I do not think the survivor's language is best understood as expressing broken quasi-trust. She has not merely had an inductive prediction about her priest or her church disappointed. Rather, I think she really has had trust in both the role-holder and the institution betrayed and her language should be taken at face value. An explanation is needed then for how institutions and role-holders can be understood as the kinds of things that can be trusted – whether reasonably or not.

Section 2.2: A Substantive Account of Trust in Institutional Contexts

Hardin's conceptual limitations claim is too strong. When trust is understood as a cluster concept it can be applied substantively to relations involving institutions and role-holders. In C*hapter 1* I argued that the concept of trust is not amenable to analysis in terms of necessary and sufficient conditions but rather should be understood in terms of a group of features that cluster together. While there may be no single feature common to all instances of trust, most cases of trust will involve certain central features. I identified these central features as: counting on someone; vulnerability to practical harm; vulnerability to moral harm; and accepting vulnerability. Instances of trust involving its central features are paradigm cases. Other relations and interactions need only be sufficiently similar to paradigm cases to be considered instances of trust themselves.

The question is whether relations involving institutions and role-holders can be sufficiently similar to exemplary cases of trust and so count as instances of trust. I think they can be. Consider the "trust fall" example I presented as a paradigm case in *Chapter 1*. In the "trust fall", one team member stands blindfolded in the centre of a circle made up of other team members. The blindfolded team member then falls backwards, counting on her teammates to catch her. This example involves practical and moral vulnerability to the teammates doing the catching and acceptance of vulnerability as one falls. It may also involve relying on the goodwill of others toward oneself, expecting others to be trust-responsive, and entrusting one's bodily safety into the care of others. That is, the blind-folded participant may entrust her bodily safety to her teammates and let herself fall because she thinks her teammates have goodwill towards her or will be moved to do their best to catch her because of the fact that she is counting on them.

Relationships involving institutions and role-holders can be sufficiently similar to the exemplary "trust fall" case. Consider again the Catholic Church example. The survivor was *vulnerable* to the church and to the priest who was the representative of that church. As she grew up as a member of the church she incrementally became vulnerable to the church and its role-holders. This is evident when she writes, "In my own early experience, the manipulation was subtle. It began when I was just seven with the demand for respect, the 'Yes, Father, no, Father' responses that we were forced to say. If we failed, we were immediately reprimanded" (Fleming, 2007, p. 143).

As I explained in *Chapter 1*, vulnerability can be explicitly and voluntarily accepted at a point of decision but it can also be implied as trust emerges incrementally over time. I think the Catholic Church example involves such implied acceptance of vulnerability. The survivor does not consider the risk of abuse and explicitly choose to accept that vulnerability; she is most likely not even aware of the risks. Rather, she enters her relationship with the Catholic Church and the priest early in life and, because this trust relationship continues to be significant in her life, she remains vulnerable to the abuse of her trust.

The survivor's continued relationship with the Catholic Church and her priest can be partially explained by the fact that she was manipulated and controlled over time. As she writes,

> Father's powerful skills of manipulation were masterfully concealed behind his potent charisma. People either acquiesced to his control and followed him or quickly sized up his motives and moved away. The vulnerable were swept into his conservative theology, charismatic presence, and subtle domination. I was one of the vulnerable ones, initiated at an early age. (Fleming, 2007, p. 143)

The survivor's continued relationship can also be explained by the quasi-perceptual affective attitude that can feature in trust. In C*hapter 1* explained Karen Jones' view of trust as an affective attitude of optimism about another's goodwill and competence. On her view, trust involves seeing another in a certain way that leads to an expectation that they will be trust-responsive – that they will be directly and favourably moved by the thought that you are counting on them (K. Jones, 1996, p. 8). I argued that Jones' view was ultimately unsatisfactory as an understanding of the concept of trust. A trickster may be optimistic about another's goodwill toward him while merely relying on him but not trusting. And because in Jones' view the expectation of trust-responsiveness is based on the attitude of optimism about another's goodwill, her view is unable to account for instances of trust where goodwill does not play a significant role. As I explained in C*hapter 1*, while Jones' view is unsatisfactory as a complete understanding of the concept of trust, features of it can be involved in trust's cluster.

The quasi-perceptual attitude that Jones identifies is helpful in understanding the Catholic Church example. From an early age the abuse survivor saw the priest who later abused her as someone with goodwill toward her; presumably, as someone who could be trusted. She writes,

> As our lives went on, Father made weekly appearances to visit, play and eat ice cream with us. He was a welcome visitor; there was happiness and laughter when he came to visit. I attended religion classes each Saturday morning and my attachment to him continued to grow. He became a second father to me, and in many ways I carried greater respect and awe for him than I did for my own father. I was too young to imagine he had a dark side. (Fleming, 2007, p. 142)

I think it is understandable how the survivor's trust in the priest and the Catholic Church he represented emerged over time as she came to see both of them positively. As her trust grew, it likely shaped the way she saw the priest and the Catholic Church as a role-holder and institution with goodwill toward her that could be expected to respond to the fact that she was counting on them. And so her interaction with them persisted and deepened as did the manipulation, control and abuse done to her. To be sure, the priest did not treat the survivor with goodwill. And in not taking responsibility for the abuse done by one of its role-holders, the Catholic Church failed to be properly responsive to the survivor's trust and the damage done to it. The church failed to be moved by the fact that the survivor was counting on it for recognition, reparation and resolution. Despite those failures, the Catholic Church case is similar to the paradigmatic "trust fall" case. Both involve someone being vulnerable to another or others; taking the other(s) to have goodwill toward them; and expecting the other(s) to be trust-responsive.

There are some differences between the paradigmatic "trust fall" case and the Catholic Church example. The vulnerability involved in the "trust fall" exercise is more easily identifiable. Blind-folded participants know that by letting themselves fall they risk falling to the ground. Their acceptance of that vulnerability is explicit when they decide to trust their teammates and fall. In contrast, the survivor's vulnerability in the Catholic Church example is established incrementally and her acceptance of vulnerability is implied. She does not see the risks and choose to interact with the Catholic Church and the priest regardless of them. Rather, she interacts with the Church and the priest taking them for what they seem to be – a caring institution and role-holder. But while the vulnerability and acceptance of it vary across these examples, the fact remains that they both involve features of vulnerability and acceptance of vulnerability. Both cases can also be understood as involving features of, at least presumed, goodwill and trust-responsiveness, though in the Catholic Church case the seeming goodwill and responsiveness on the part of the priest is a facade used for manipulation. As I explained in C*hapter 1*, instances of trust need not be exactly like paradigm cases of trust. That should not be expected from a cluster concept that has no single feature, or features, common to all its instances. Rather, relations need only be sufficiently similar to paradigm cases of trust in order to count as instances of trust themselves. I think the fact that the Catholic Church example involved vulnerability, acceptance of vulnerability, goodwill and trust-responsiveness makes it sufficiently similar to the paradigmatic "trust fall" case and so it can be taken to be an instance of trust. As such, I take the Catholic Church example to be a counterexample to Hardin's conceptual limitations claim. The concept of trust *can* be applied to relationships involving individuals, institutions and role-holders.

It might be thought that the Catholic Church example I have used to argue that the concept of trust is applicable to the institutional context is somehow substantively different from the government institutions Hardin has in mind. There may be some reason to think that relations involving *religious* institutions are sufficiently similar to paradigm cases of trust while those involving *government* institutions are not. Perhaps persons are not vulnerable to government institutions, do not accept vulnerability to them, or have less of a background expectation that they and their role-holders will care for them. But I do not think this is true. Relations involving Government institutions and role-holders can also be sufficiently similar to paradigm cases of trust. Consider the case of children emigrating from the British Isles to Australia in the post World War II era. After briefly presenting the story of the child migrants I will show that they – and their parents where present – were vulnerable to government institutions and role-holders, either explicitly or implicitly accepted vulnerability, and had a background expectation that they – or in the case of the parents, their children – would be cared for by such institutions and role-holders.

In 1944 the Commonwealth introduced a scheme to introduce 50,000 child migrants into Australia, the majority of whom would be "war orphans"[35] (Gill, 1997, p. 55). Many of

[35] Australia was not the only destination to which child migrants were sent. They were also sent to colonies in

the children had been living in state run “homes”[36] in the UK after the deaths of their parents while others were introduced to the Commonwealth scheme by parents who, for various reasons, considered themselves to be unable to care for their children. However the children entered the scheme, emigration was supposedly voluntary with any parents or guardians involved giving their permission and the children signing consent forms before boarding ships bound for Australia (Gill, 1997). But in reality, many children were sent from “care” facilities in Britain to Australia without their parent’s or guardian’s knowledge or consent (Gill, 1997, p. 8). The “informed consent” given by the children is also dubious. Gill tells the story of one boy who upon reading a notice stating that, “‘The next party for Australia will leave on 30 March’ thought he was putting his name down for a party” (Gill, 1997, p. 93).

Upon arrival in Australia, child migrants were sent to receiving institutions that were government sponsored and in most cases Christian charities (Gill, 1997, p. 3). Gill explains that such institutions were drawn to the migration scheme in part because the Commonwealth government and Australian state governments agreed to fund one-third of the costs of institutional buildings for housing child migrants, accommodation for institutional staff, and additional furniture and equipment.

> Correspondence between ecclesiastical bodies mentions the benefits that would accrue. Put crudely, it would appear that the religious Orders hoped to get new buildings (and in the case of the Christian Brothers, the children themselves as labourers), which would serve their long term interests, in return for the inconvenience of looking after British child migrants. (Gill, 1997, p. 62)

Having been conned into “voluntarily” emigrating to an unfamiliar environment, isolated from the familiar, given poor food and lodging, and often made to work long hours, child migrants were extremely *vulnerable* to their “care” givers. Not only were they vulnerable to

Canada, Rhodesia (Zimbabwe), and South Africa. (Gill, 1997, p. 3) In the 1960’s British and Australian governments began to lose interest in the Commonwealth child migration scheme. The ending of the scheme was not as a result of significant child welfare reform. Rather, it seems to have faded away quietly (Gill, 1997, p. 61).

[36] I follow Joanna Penglase in writing “care” and “homes” in quotes so as not to confuse the experiences children had in state institutions with caring or the places they inhabited with places of refuge in any real sense (Penglase, 2007, p. 39).

institutions and role-holders because of their age and isolation, they were vulnerable because they were in a new country with new customs and rules and in a new environment that would have been very unfamiliar to them[37]. Those parents who entrusted their children to institutional "care" were also vulnerable to the institutions' and role-holders' "care" of their children. Their vulnerability is highlighted by the fact that they were lied to about the "care" of their children and did not even know that their children had emigrated to Australia.

I think that, as in the Catholic Church example, acceptance of vulnerability is for the most part implied in the child migrant case. While the children had supposedly "voluntarily" signed up to emigrate, many did not know what they were signing up for; and they could not have known all the risks of doing so. They were simply not given that information. Rather than knowing the risks involved in emigrating to Australia and then voluntarily deciding to accept those risks like the blind-folded participant in the paradigmatic "trust fall" case, the child migrants seem to have simply gone along with their "care" givers. In contrast, some of the parents entrusting their children to state-run "care" institutions could be understood to have explicitly accepted vulnerability to institutions and role-holders. They might not have been aware of the risk of being lied to about the whereabouts of their child, but they would have presumably known that they were, at some point, deciding to leave their children in the discretionary care of others.

The child migrant case can also be understood as involving a background normative expectation of goodwill and trust-responsiveness. The institutions involved in the child migrant schemes and their role-holders, had a duty of care to the children entrusted to their "care" and ought to have been responsive to this duty and to the fact that the children were vulnerable to them and had little recourse but to count on them. Presumably parents who entrusted their children to institutions assumed that institutions and their role-holders would

[37] The significance of the harsh Australian environment is highlighted in *Oranges and Sunshine,* a recent film portraying the child migrant experience (Loach, 2010).

be mindful of this duty and so assumed that their children would be well cared for. The child migrants themselves also presumably expected that their "care" givers would care for them, since many seemed to have believed what their "care" givers told them about Australia being "a land of sea and sunshine [where] children rode on Kangaroos' backs" (Gill, 1997, p. 8). By failing to treat the children with goodwill and to be appropriately responsive to their vulnerability, the institutions and role-holders "caring" for the child migrants did not uphold the normative expectations implicit in their duty of care – that is, to be sensitive to the children's physical and emotional needs and, given those needs, to provide appropriate care.

Given the features of vulnerability, (explicit and implicit) acceptance of vulnerability, and expectation of goodwill and trust-responsiveness involved in the child migrant case I think it is, like the Catholic Church case, sufficiently similar to paradigm cases of trust. That relationships between child migrants and the government institutions in charge of their "care" involved trust is further supported by the fact that apologies have been made to surviving child migrants on behalf of the Australian Federal Government and the British Government. In November 2009 the Australian Federal Government acknowledged the abuse done in government supported homes (Rudd, 2009). And in 2010 then British Prime Minister Gordon Brown made a formal apology for the child migration scheme (Brown, 2010).[38] In his apology, then Prime Minister Kevin Rudd said,

> Sorry – that as children you were taken from your families and placed in institutions where so often you were abused. Sorry – for the physical suffering, the emotional starvation and the cold absence of love, of tenderness, of care. Sorry – for the tragedy, the absolute tragedy, of childhoods lost,– childhoods spent instead in austere and authoritarian places, where names were replaced by numbers, spontaneous play by regimented routine, the joy of learning by the repetitive drudgery of menial work. Sorry – for all these injustices to you, as children, who were placed in our care [...] We look back with shame that many [of] these little ones who were entrusted to institutions and foster homes instead, were abused physically, humiliated

[38] In March 2001 the Roman Catholic Church in Australia also made a formal apology to child migrants during a senate enquiry (Dutter, 2001). But as I am concerned with showing that the concept of trust applies to relations involving government institutions, I focus on apologies by government leaders on behalf of the government institutions involved in the migration scheme.

> cruelly, violated sexually. And we look back with shame at how those with power were allowed to abuse those who had none. [...] The institutions the nation created for your care, failed you...A turning point for governments at all levels and of every political hue and colour to do all in our power to never let this happen again. For the protection of children is the sacred duty of us all...Because let us be clear - these children, both from home and abroad, were placed in care under the auspices of the state, validated by the laws of the land... (Rudd, 2009)

Contra Hardin's conceptual limitations claim, I take the above child migrant case to show that the concept of trust can be applicable to interactions involving government institutions and role-holders, as demonstrated by Kevin Rudd's apology. Nevertheless, Hardin's epistemic limitations claim may still stand. Trusters might not be able to know enough about institutions and their role-holders to have good reason to trust them. In the following section I argue that Hardin is not correct about this and that individuals *can* know enough to have good reason to trust institutions and role-holders.

Section 3: Trustability in Institutional Contexts

The concept of trustability captures the attribute of being one whom another has good reason to trust. In C*hapter 2* I argued that persons might be trustable because they are competent and committed (because of various motives) or have trustworthy, virtuous character. In this section I argue that individuals can have trust-relevant knowledge in institutional contexts. Specifically, I show that individuals can at least know enough to have reason *not* to trust institutions and role-holders; and that individuals can know enough to have good reason *to* trust in institutional contexts. That is, institutions and role-holders can be trustable.

Individuals can at least know enough *not* to trust institutions and role holders. This is evident in the social environment of the German Democratic Republic (GDR). In *Stasiland*, Anna Funder depicts the climate of distrust present in the GDR: "Relations between people were conditioned by the fact that one or other of you could be one of *them*. Everyone suspected everyone else, and the mistrust this bred was the foundation of social existence" (Funder, 2002, p. 28).

Despite its name, the German Democratic Republic lacked all semblance of democracy. Instead, an extreme power imbalance existed between citizens and the strong arm of the government – the Stasi officers. Funder writes,

> There were, at least on paper, political parties other than the ruling Socialist Unity Party. But really there was just the Party, and its instrument, the Stasi. Judges often got their instructions from the Stasi which in turn, passed them on from the Party – right down to the outcome of judgement and length of the sentence. The connection of the Party, the Stasi and law went from the ground up: the Stasi, in consultation with school principals, recruited obedient students with an appropriately loyal attitude for the study of law. I once saw a list of dissertation topics from the Stasi Law School at Potsdam, which included such memorable contributions to the sum of human knowledge as 'On the Probable Causes of the Psychological Pathology of the Desire to Commit Border Infractions'. There was no room for a person to defend themselves against the State because all the defence lawyers and all the judges were part of it. (Funder, 2002, p. 37)

The GDR was a smoothly operating surveillance machine. In the end there were 97,000 Stasi employees and over 173,000 other informers spread throughout the population (Funder, 2002, p. 57). But, as is shown by the stories of those Funder speaks with, citizens of the GDR did not live in a climate of distrust just because of the number of surveillance officers and informants. They also inhabited that climate because of their knowledge that anyone could be an informer. One woman tells Funder, "'I conformed, just like everybody else. But it's not true to say the GDR was a nation of seventeen million informers. They were only two in a hundred'" (Funder, 2002, p. 74). Funder continues, "Even with one informer for every fifty people, the Stasi had the whole population covered" (Funder, 2002, p. 74). I take the citizens in the GDR to have had good reason not to trust their government and its role-holders. All that was needed was to know that persons in their community, perhaps their neighbour, their doctor or their postal worker, were informants. They knew enough to have reason to think such parties were untrustable.

I take the above example from *Stasiland* to show that individuals can have trust-relevant knowledge in institutional contexts. But persons can also know enough to have good reason to trust institutions and role-holders. As in inter-personal relationships, individuals need only have good reason to think an institution or a role-holder is competent and committed to doing what he is counting on them to do to for his trust in them to be reasonable. For example, if a citizen has good reason to think her local police force is competent at preventing and dealing with crime in her neighbourhood and committed to protecting individuals like her from criminals, she may have good reason to trust the police force. However, individuals may have good reason to think that an institution or role-holder is competent and committed to do doing what they are counting on them to do only to have that institution or role-holder break their trust. For example, given her early experience, I think it was reasonable for the survivor in the Catholic Church case to trust the church and her priest. For all she knew, they were competent and committed to caring for her. But in actuality she was being manipulated and controlled. Given the seeming goodwill and care the priest showed her as a child, the survivor had reason to see the church and the priest as trustable; but they were not. And so the question is not just whether individuals can have good reason to think an institution or role-holder is trustable but whether it is possible for individuals to know if institutions and role-holders *really are* competent and committed to doing what they are counting on them to do.

O’Neill has identified two mechanisms for grounding trust: what she calls, “intelligent accountability” and “active checking” (O'Neill, 2002b, pp. 57, 77). I will show that policies making institutions and role-holders accountable in the way O’Neill has in mind can encourage them to be competent and committed to doing what individuals are counting on them to do. While such accountability does not guarantee that an institution or role-holder is in fact trustable, it can support their being trustable. Because accountability does not guarantee trustability, I consider if institutional transparency might reveal whether institutions and role-holders really are trustable. But I follow O’Neill in holding that mere transparency is susceptible to deception. And we do not just need to know more about institutions and role-holders; we need to be able to confirm that the information we receive about their competency and commitment is true. That is, we need to be able to actively check that information.

Not all types of accountability will encourage trustability in institutional contexts. In developing her view of “intelligent accountability”, O’Neill contrasts it with a different type of accountability focused on control which, in fact, can hinder the competence and commitment of institutions and role-holders to do what individuals are counting on them to do. This type of accountability, which she calls “new accountability” (O'Neill, 2002b, p. 47), uses standardized performance indicators and extensive documentation to monitor the behaviour of those being held to account.

> Detailed instructions regulate and prescribe the work and performance of health trusts and schools, of universities and research councils, of the police force and of social workers. And beyond the public sector, increasingly detailed legislative and regularity requirements also bear on companies and the voluntary sector, on self-employed professionals and tradesmen. All institutions face new standards of recommended accounting practice, more detailed health and safety requirements, increasingly complex employment and pensions legislation, more exacting provisions for ensuring non-discrimination and, of course, proliferating complaint procedures. (O'Neill, 2002b, pp. 46-47)

“New accountability” can fail to encourage trustability by obscuring the things an institution or role-holder is being counted on to do. O’Neill points out that performance indicators are often chosen for ease of measurement rather than because they accurately identify quality completion of the specific aims of a given institutional practice. For example, I take it that police departments and their officers are counted on by the public to work to reduce crime in the community and enhance public safety. That a police department or police officer clears up crimes quickly does not necessarily mean the department or officer is competent or committed to reducing crime and enhancing public safety. Rather, dealing with crimes quickly may suggest incompetence and a lack of commitment if quality and care in policing is sacrificed. To have good reason to trust police departments and officers, individuals presumably do not just want to know that crime is dealt with quickly but that it is dealt with well and in a way that aids in the safety and well-being of their communities.

In addition to obscuring what institutions and role-holders need to be held accountable for, "new accountability" can fail to encourage trustability insofar as it hinders role-holders from competently doing what they are being counted on to do. O'Neill explains,

> Professionals have to work to ever more exacting – if changing – standards of good practice and due process, to meet relentless demands to record and report, and they are subject to regular ranking and restructuring. I think that many public sector professionals find that the new demands damage their real work. (O'Neill, 2002b, p. 49)

For example, the ability of police to reduce crime and enhance public safety may be decreased because of extended documentation required to deal with each case they take on.

In contrast to "new accountability" which attempts to micro-manage and control those being held accountable, "intelligent accountability" involves attention to institutions and role-holders governing themselves well given their particular tasks and functions (O'Neill, 2002b, p. 58). For example, rather than merely holding a police officer to account for the speed and efficiency with which she deals with a crime, on O'Neill's view she would be held to account for the way in which she handled a given crime in its particularity.

Given its focus on particularity, "intelligent accountability" is not best served by standardization. For example the Centre on Housing Rights and Evictions (COHRE) explains that in preparation for the 1996 Olympic Games in Atlanta, Georgia 9,000 arrest citations were given to the homeless in the community as part of a campaign to "clean the streets" (COHRE, 2010). Some non-homeless members of the community might have viewed the actions taken by the Atlanta police as reducing crime and enhancing public safety in their community. But the actions taken by the police did not satisfy their role with regard to the homeless members of the community. And so holding the police to account for some standardized indication of "reducing crime and enhancing safety in the community" without considering the particularities of public safety needs would not necessarily encourage the police to be competent and committed to doing what individuals are counting on them to do.

Rather than meeting standardized performance indicators, O'Neill explains that "intelligent accountability" involves giving an account to a third-party that is able to satisfactorily assess and report on how an institution or role-holder has governed themselves (O'Neill, 2002b, p. 58). The focus is on assessing the ways in which institutions and role-holders govern themselves given specific situations and on a case by case basis rather than merely requiring them to meet a standardized requirement. For example, rather than having to meet a quota for dealing with a certain number of crimes, a police department and its officers could be held to account to an independent organization consisting of individuals with sufficient time and experience to assess and report on the functioning of the department and its officers.

Policies putting "intelligent accountability" in place can encourage competence and commitment in institutional contexts. Where institutions and role-holders are made to account for the way they govern themselves incompetence can be identified and addressed. For example, a police department may establish and implement a policy requiring the department to allow periodic observation of its policing by an external human rights organization. Given their observations, the external organization could inform the police department of any incompetence it has identified so that those failings can be addressed. Such a policy would not attempt to control the behaviour of the police department or its officers but would rather allow for them to govern themselves and have that governance assessed. Further, if "intelligent accountability" is established, role-holders are more likely to be motivated to do what they are supposed to do given their roles. This is because role-holders will be aware of the fact that they will actually be held responsible for appropriate conduct given their roles. As long as individuals are counting on role-holders to act in ways consistent with their roles, such accountability can support role-holder trustability. That is, by being motivated to fulfil their role, role-holders would be in effect being motivated to do what individuals are counting on them to do. For example, if a doctor knows she will be held to account for making accurate diagnoses of patient conditions and if a patient is counting on the doctor to make such diagnoses, then the doctor's motives to have a positive outcome in terms of accountability also make her committed to doing what the patient is counting on her to do.

While policies implementing "intelligent accountability" can encourage trustability in institutional contexts they cannot establish that trustability. That is, they do not necessarily make an institution or role-holder competent and committed to doing what one is counting on them to do. Despite policies for "intelligent accountability" being in place, institutions and role-holders may fail to comply with them. Or they may seem to comply but do so deceptively. For example, even if the abusive priest in the Catholic Church case had to account regularly for his actions, he might have simply deceived those assessing his conduct and his destructive behaviour might have continued without being discovered.

It might be thought that increasing transparency aids trust in institutional contexts; that is, the more information that is available to individuals about institutions and role-holders increases their ability to assess whether those institutions and role-holders are actually competent and committed. But, as O'Neill points out, mere transparency is not enough because secrecy is not the enemy of trust. Deception is (O'Neill, 2002b, pp. 70-72). And while increasing transparency makes more information available, transparency does not minimize deception.

> ...unless the individuals and institutions who sort, process and assess information are themselves already trusted, there is little reason to think that transparency and openness are going to increase trust. Transparency can encourage people to be less

> honest, so increasing deception and reducing reasons for trust: those who know that everything they say or write is to be made public may massage the truth. (O'Neill, 2002b, p. 73)

Because transparency does not necessarily minimize deception, and can even encourage it, individuals do not just need to know *more* about an institution or a role-holder; they need to know whether the information they receive about an institution's or role-holder's competence and commitment is true. And to know that the information available to them about the institution and its role holders is true, individuals must know that the source of their information is honest and reliable; they must be able to actively check that information. As O'Neill writes, "Where we can check the information we receive, and when we can go back to those who put it into circulation, we may gain confidence about placing or refusing trust" (O'Neill, 2002b, p. 76). O'Neill explains that "active checking" is not just confirming information across several sources or confirming it with one's favourite source of information. "...Arguments from authority, to use the old term, however deliciously congruent with favourite beliefs, establish nothing" (O'Neill, 2002b, p. 77). Rather, "active checking" is simply checking information for oneself as best as one can. O'Neill cites informed consent as an instance of "active checking" (O'Neill, 2002b, p. 78). Informed consent is essentially agreeing to something after having considered the relevant information and risks involved. But informed consent only works if persons deciding whether to give their consent are able to further check the information given them. So too with trust in institutional contexts. To know whether an institution or role-holder really is trustable – is competent and committed to doing what they are being counted on to do – individuals will need to be able to check the available information. And they will need to be able to further check the source of that information.

It may be thought that "active checking" is contrary to the concept of trust. As I explained in C*hapter 1*, entrusting some valued good to the discretionary care of another involves not checking up on their care of it. While I do not think all trust involves entrusting, I do think that checking up on another’s discretionary care is contrary to the kind of trust that does involve entrusting. For example, in the paradigmatic “trust fall” case the blind-folded participant entrusts her physical safety, if only for a moment, to the discretionary care of her teammates. They may choose to catch her; they may choose to let her fall; and they may choose to catch her in a number of ways. If the blind-folded participant were to take off her blind-fold and fall while checking to see how her teammates were going to catch her, I do not think it would be correct to say that she was entrusting herself to their discretionary care. She still might be counting on them to catch her but she is not entrusting her physical safety to them.

In contrast, I do not think the “active checking” involved in institutional contexts is contrary to counting on an institution’s or role-holder’s discretionary care. Instead, “active checking” is a matter of checking information *before* one trusts an institution or role-holder. It is involved in coming to see an institution or role-holder as trustable and then trusting them because they are competent and committed to doing what you are counting on them to do. And that need not be contrary to trust – even in interpersonal relationships. For example, the blind-folded participant in the “trust fall” may pause before letting herself fall and may say to her teammates, “You are going to catch me right?” Then, after being encouraged by their resounding “Yes!” the blind-folded participant may accept vulnerability to her teammates and fall into their arms. She is still entrusting her physical safety to the discretionary care of the teammates but she is only doing so after checking up on their commitment to catching her. Similarly, actively checking information about institutions and role-holders need not be contrary to trust in institutional contexts. It is, rather, a way of coming to know whether an institution or role-holder is competent and committed to doing what one is counting on them to do and then, given what one discovers, trusting them or not.

As O’Neill has pointed out, “active checking” can be demanding. It is unlikely that individuals will always be able to trace information about institutions and role-holders all the way back to their original sources to check its truthfulness. But where institutions have made the sources of such information about themselves available, it is plausible that individuals will be able to actively check information supporting or undermining the trustability of those institutions and/or their role-holders.

Merely being able to actively check information about an institution or role-holder does not mean they will necessarily be trustable. Upon checking, one may find out that an institution or role-holder really is *in*competent and *un*committed to doing that which one is counting on them to do. For example, recall the case of the child migrants I presented in section 2.2. As I explained, some parents who entrusted their children to institutional “homes” were deliberately lied to and their consent to have their children immigrate to Australia was forged (Gill, 1997, p. 8). If the parents who had entrusted their children to the institutions had checked the information they received about the whereabouts of their children, they would have found that the “carers” to whom they had entrusted them were in fact not to be trusted. So while “active checking” can be a useful tool for testing information about an institution or role-holder, that an individual can actively check information about institutions and role-holders does not establish trustability. Rather, for institutions and role-holders to be trustable there must be policies in place that encourage them to be competent and committed to doing what an individual is counting on them to do and they must follow through with those policies showing that they are actually competent and committed in that way.

The policies, and compliance with them, that suffice to make institutions and role-holders trustable to some individuals will not necessarily make them trustable to others. When there is a troubled history between individuals, institutions, and role-holders, requirements for establishing and expressing competence and commitment to another will be more demanding than if that history was not present. Consider again the apology given by Mr. Rudd on behalf of the Australian Federal Government to child migrants (Rudd, 2009). Australian citizens who were not affected by the abuse in government "homes" might think that Mr. Rudd's apology shows him, and the government he leads, to be committed to repairing broken trust. They may think that the government is finally acknowledging the wrongs that were committed against child migrants in Australia. But to the ears of one of the surviving child migrants Mr. Rudd's words may not have that effect. They may take the apology to be mere "lip service" that does not really change their view of the government. Instead, survivors may think that a federal reparations policy for wrongs done to child migrants is required to show that Rudd and his government are committed to dealing with the broken trust between child migrants and the Australian Federal Government.

That some institutional policies, and compliance with them, can make an institution trustable to some but not others shows that trustability is context sensitive. It is not a matter of whether one is competent and committed in some general sense, but whether one is competent and committed to doing what a specific individual is counting on one to do. In addition to trustability being context sensitive, coming to see another as trustable is *stance sensitive*: it is dependent on how one perceives another. And that perception can be affected by various things. As in the child migrant example above, whether someone takes another to be trustable can be affected by the truster's relational history with that person. But it can also be affected by what Annette Baier has called "climates of trust" (Baier, 1986, p. 245). Baier introduces the idea of climates of trust in discussing the way in which trust can be based on a kind of presumption that role-holders will comply with the policies shaping their roles. Climates of trust can arise because of the fact that an institutional role is well established in a community. Persons may just assume that role-holders will properly perform their roles. As Baier writes,

> We take it for granted that people will perform their role-related duties and trust any individual worker to look after whatever her job requires her to. The very existence of that job as a standard occupation, creates a climate of some trust in those with that job. (Baier, 1986, p. 245)

It is plausible that the abuse survivor in the Catholic Church example had such a presumption about the priest who ended up abusing her. In his role in the Catholic Church, the priest was meant to be a representative of Jesus Christ; a go-between for parishioners and God. Presumably, that role brought with it expectations of a high level of morality and care for the vulnerable. The survivor herself says she even thought the priest and other clergy were “without human flaw” (Fleming, 2007, p. 143).

Once a climate of trust has been established persons may perceive an institution to be trustable even if they do not know much about the roles or policies that were significant in establishing that trustability in the first place. For example, I presume Australia Post to be trustable. I express this in the way I confidently entrust the care of parcels to them. When I do so I have little concern for whether postal workers will violate my trust, open my parcels and help themselves to whatever contents they may find. Rather, I assume they will not do so. I do not know much about Australia Post’s privacy policies, its worker conduct policies or its role-holders. But my previous experiences with the *U.S.* Postal Service were positive. And upon arrival in Australia I saw no reason to think Australia Post would conduct itself with any less integrity. Rather I unreflectively assumed Australia Post and its role-holders to be competent and committed to handling my mail well. Further, that assumption was not called into question by the behaviour of persons around me. I do not see Australians treating postal workers with suspicion as they leave parcels at the post office. Instead, mostly without thinking about it, I enter a climate of trust regarding Australia Post that was established long before my arrival in Australia.

Because of climates of trust persons do not come to every interaction with an institution or role-holder with a "clean slate", so to speak. Rather, we come inhabiting an environment shaped both by our past experiences and by the trust, or distrust, of others who have interacted, and are interacting now, with a given institution. If the climate we are in is one of trust we will be more likely to take ourselves as having reason to trust. Conversely, if that climate is one of distrust, we will be less likely to do so. I will return to explicate distrust and its relation to trust in C*hapter 5* when identifying the ways that betrayal can damage trust.

The impact that climates of trust can have on our perception of an institution's or role-holder's trustability means that we will need to be careful when accepting vulnerability to them because we think they are trustable. Our perception of them may be influenced by a presumption that they are competent and committed to doing that which we are counting on them to do. This is where "active checking" can be helpful. We can check to see if an institution or role-holder really is competent and committed to doing what we are counting on them to do. But given climates of trust, we should not stop at checking information about institutions and their role-holders. We should also critically assess our own assumptions about those institutions. Climates of trust can be established for good reason; most of the time Australia Post does prove itself to be competent and committed to delivering my post well. But it can still be prudent to critically assess the climates of trust we inhabit. Those which are there for good reason should be able to stand up to such assessment.

In this section I have shown that it is possible to know enough about institutions and role-holders for them to be trustable. While trusting institutions involves the risk that they will let us down, manipulate us and even betray us, trust in institutional contexts can be reasonable. As I explained in *Chapter 2,* that a trustee is trustable does not mean he is trustworthy. Recall the example in C*hapter 2* of the reliable and even trustable, but untrustworthy hit man. The hit man is skilled in his role in the mob. He always makes his hits. The mob boss knows the hit man is competent to do the jobs he gives him and that he is committed to doing them. The hit man is a reliable killer and, to the mob boss at least, he is trustable. But it would not be correct to call the hit man trustworthy. As I argued in *Chapter 2*, the trustworthy are competent and committed but that is not all; they have a kind of integrated, virtuous character disposing them to act in ways which contribute to human flourishing. It is because they have such character that the trustworthy respond appropriately to those counting on them. The hit man's murderous activities reveal that he lacks such character. In the next section I explain how this character-based view of trustworthiness is applicable to institutional contexts.

Section 4: Trustworthiness in Institutional Contexts

The character-based view of trustworthiness I argued for in *Chapter 2* can be applied to institutional contexts, although applying it to institutions and role-holders faces two challenges. First, it is unclear whether institutions can be understood to have character at all. Second, even if institutions can have character, an explanation of how that character can be trustworthy or virtuous is required. In section 4.1 I show that institutions can have character. Then I develop a substantive account of integrated, virtuous character in institutional contexts. In section 4.2 I discuss Oakley and Cocking's view of virtuous professional roles. Then in section 4.3 I apply their view and the moderate list of requirements for trustworthiness I adapted from Potter in *Chapter 2* to institutions and role-holders. I argue that institutional trustworthiness depends on a complete package of institutional functions, roles and policies which dispose an institution and its role-holders to contribute to human flourishing.

Section 4.1: An Account of Institutional Character

Character is of central importance to trustworthiness. As I argued in *Chapter 2,* trustworthiness marks a kind of integrated, virtuous character, which disposes persons to contribute to human flourishing. However, it is unclear whether institutions have character. Rather, it might be that talk about an institution being trustworthy just means that the individual role-holders within the institution are trustworthy. This would be to take a summative approach and to understand an institution's character as an agglomeration of its role-holders' character. For example, on a summative approach saying that Royal North Shore Hospital has virtuous character is equivalent to saying that its doctors, nurses and other representatives have virtuous character.

Taking a summative approach to institutional character is problematic because an institution can have traits that differ from those of its members. Consider Reza Lahroodi's example of a church committee, which as a group is prejudiced against gay rights despite its members resisting such prejudice personally.

> As individuals, all or most members of the committee routinely resist their initial tendency to dismiss ideas favouring gay rights that are contrary to their own and grant them enough plausibility to take them seriously. The group, however, moves in the opposite direction. It fails to assign any plausibility to a wide range of contrary views about gay rights, summarily dismisses them and does not consider them worthy of discussion, let alone adoption. [...] In this way, a group may be locally narrow-minded, while all or most of its members are locally open-minded. (Lahroodi, 2007, p. 287)

Lahroodi's example concerns a committee, but the point applies to various types of groups and institutions. At issue is the interplay between individual role-holders and the larger entity to which they belong. Because, as Lahroodi's example shows, it is possible for a group to fail to reflect the views of its members, groups should not just be understood as the sum of their members; and the character of institutions should not just be understood as the sum of their role-holders' character.

Todd Jones presents a non-summative view of groups that provides a plausible start to an account of institutional character. I say that T. Jones'[39] view provides a *start* to an account of institutional character because it shows that institutions can have character traits, but as I will show, it needs to be supplemented with an explanation of what determines the traits an institution has and expresses.

On T. Jones' view, functions of groups can be identified, and traits can be attributed to groups based on the way they fulfil their functions. On his functionalist view, what something does is more important than what it is made of. Just as a computer has many parts working to solve equations, a group can have many members working to complete a given job. And just as traits can be applied to computers with regard to the efficiency with which they solve equations – i.e. "this computer solves equations *quickly* and *reliably*" – traits of a group can be identified. Like a computer, a group of mathematicians may be organized to solve equations quickly and reliably. Or, as T. Jones writes, "A group of people, for example, could be organized to be a reliable surveillance system, collecting and passing on data for government organizations on where weapons caches are hidden..." (T. Jones, 2007, p. 441). In T. Jones' example, the function of the group is to be a surveillance system and the characteristic, or trait, applied to the group is that of being reliable.

Understanding groups in terms of their function might run into a homunculus problem if the "little person" that is posited as the agent *within* the group cannot be found. But in a group of persons the homunculi are ready made and found in the presence of individual group members. As T. Jones puts it,

> Now in the philosophy of mind, the explanation of human cognitive processes by little men inside one's head is meant to be purely metaphorical. But if one were to take this metaphorical story completely literally and use it to describe, not the thinking of individual human agents, but the activities of groups, one could give an account of the

[39] Throughout the rest of this chapter I will use 'T. Jones' to differentiate the work of Todd Jones from that of Karen Jones, whose work I have discussed at length in C*hapters 1* and *2*. Because I discuss her work more often throughout this thesis, I will simply use 'Jones' when referring to the work of Karen Jones in this chapter.

> unitary cognitive activities of a complex agent constructed out of the micro-activities of individual group members. (T. Jones, 2007, p. 441)

T. Jones' explanation of his functional view in terms of group member activities does not reduce his view to a summative approach. That is, saying that a group functions as an agent made up of how individual members act does not necessarily mean that it is merely the sum of its members or their actions. On the contrary, the point of T. Jones' functionalism is that a group can function in a way that is *beyond* the sum of its members and their actions:

> One might similarly find that each member of a group of police detectives is not very tenacious in investigating a particular homicide on his own, but put them together and the dynamics of the group could make it the case that the group as a whole comes to have the epistemic virtue of being very tenacious in its pursuit of the truth. (T. Jones, 2007, p. 446)

Understood functionally then, groups can have substantive traits that differ from those of their members.

Institutions can also be understood functionally. For example, consider the institution of The Australian Department of Immigration and Citizenship (DIAC). It might be considered summatively such that DIAC just is the sum total of its immigration agents. And in one sense that is true; DIAC *is* made up of individual personnel. But DIAC and its agents can have different traits. DIAC agents may be friendly, helpful and accommodating without DIAC having those traits as an institution[40]. DIAC's requirements for entry into Australia may not necessarily be friendly, helpful, or accommodating. Nor are they supposed to be. Presumably, one of DIAC's functions is to scrutinize applicants for entry into Australia in order to guard against fraudulent migration. So while it is made up of role-holders, DIAC is not best characterised by the sum total of their characteristics. But traits can be ascribed to DIAC in functional terms. It can be taken to serve as a system for migration control, border

[40] It is unlikely that every migration applicant would describe all DIAC personnel as friendly, helpful and accommodating. Applicants from some countries may be subjected to racial discrimination and receive less friendly, helpful, and accommodating responses from DIAC personnel than other applicants. But the usefulness of my example does not require that all persons think all DIAC personnel are friendly, helpful, and accommodating. Rather, only one person need think that the DIAC personnel she deals with has such traits while recognizing that DIAC as an institution does not have them. And that, I think, is a plausible scenario.

security or whatever the goal of DIAC is taken to be. We can then attribute traits to DIAC based on the ways in which it carries out its functions. DIAC may be a *reliable* system for migration control. And among other things, it may be an *unaccommodating* but *disciplined* system for border security.

Character traits can be attributed to institutions based on the way they fulfil their functions. But saying that an institution like DIAC is a reliable system for migration control or an unaccommodating, or disciplined system for border security fails to identify the factors that contribute to its character. That is, the functionalist approach I have just explained shows that institutions can have character but it does not explain what makes them have the character that they do; it does not identify what disposes them to be reliable, unaccommodating, disciplined, efficient, and tenacious or whatever other traits they have.

What contributes to DIAC being reliable, unaccommodating and disciplined? These traits can be explained by considering the roles and policies established in DIAC. Although institutional character is not just the sum of the character of the role-holders in an institution, the roles that an institution puts in place do affect whether it fulfils – or fails to fulfil – its functions and the way in which it does so. For example, the role of an immigration officer in DIAC is presumably to scrutinize applications for migration into Australia, admit those that meet relevant requirements, deny those that fail to meet the requirements, and report those that appear fraudulent. These features of an immigration officer's role directly contribute to DIAC fulfilling its functions of scrutinizing applicants for entry into Australia and guarding against fraudulent migration. If that role had not been created in DIAC it would not fulfil its functions as reliably. And it would not be as unaccommodating to those applying for migration into Australia.

The policies established in DIAC can also contribute to the way it fulfils its functions. For example, policies requiring multiple immigration officers to review possibly fraudulent migration applications can contribute to DIAC being a reliable system for migration control. Or a policy requiring visa assistance telephone operators to withhold specific information regarding the time it will take for a given application to be processed, contributes to DIAC fulfilling its function in an unaccommodating way. Along with the roles established in DIAC, these policies serve to dispose the institution to fulfil its functions in certain ways.

Given the contribution that roles and policies can make to the way an institution fulfils its functions, institutional character is best understood as consisting of an institution's functions, the way it fulfils its functions, and the roles and policies disposing it to fulfil its functions in the ways it does. On this supplemented functionalist account of institutional character, the character-based account of trustworthiness I argued for in C*hapter 2* is applicable to institutions. Trustworthy institutions will be those with functions, traits determined by the way they fulfil their functions, and roles and policies consistent with the integrated, virtuous character that characterises trustworthiness. But exactly how such character should be understood in institutional contexts requires explanation. I will develop a substantive account of integrated, virtuous character for institutions by using Oakley and Cocking's view of virtuous professional roles and the requirements for trustworthiness I adapted from Potter in C*hapter 2*. I now turn to explicate Oakley and Cocking's view before employing it in conjunction with my moderate version of Potter's list.

Section 4.2: Virtuous Professional Roles

Oakley and Cocking have developed a substantive virtue ethics approach to good professional roles that can be extended to explicate how institutions can have integrated virtuous character (Oakley & Cocking, 2001).[41] Using an Aristotelian approach to virtue

[41] Oakley and Cocking aim to show how virtue ethics can provide an alternative to utilitarian and Kantian understandings of professional roles (Oakley & Cocking, 2001, p. 1). I am more concerned with the virtue theory account they develop with regard to professional roles than with the ways in which it is preferable to alternative approaches to morality. So I leave to one side their arguments in favour of virtue ethics over utilitarian and Kantian approaches.

ethics, they explain good professional conduct, good professional roles, and good professions in terms of how they contribute to human flourishing. As I will explain, good conduct is conduct consistent with a good role, a good role is one which aids in the fulfilment of the ends and functions of a good profession, and a good profession is one with functions which contribute to human flourishing. I will outline this view after briefly explicating the Aristotelian approach to virtue ethics and the concept of "regulative ideals" that grounds it (Oakley & Cocking, 2001, pp. 1, 25-31).

The approach to virtue ethics taken by Oakley and Cocking is consistent with the Aristotelian approach I adapted from Potter in *Chapter 2* insofar as they understand virtuous character in terms of its contribution to human flourishing. As they write,

> In this book, we base our arguments on the Aristotelian approach to grounding the character of the virtuous agent, and we take the eudemonistic view that the virtues are character-traits which we need to live humanly flourishing lives. (Oakley & Cocking, 2001, p. 18)

Oakley and Cocking explain the relation between an ethical theory's "criterion of rightness" (Oakley & Cocking, 2001, p. 1) and the way agents are guided by that criterion in terms of the notion of a "regulative ideal" (Oakley & Cocking, 2001, pp. 1, 25-31). A regulative ideal is a conception of correctness or excellence that, when internalised, influences one's motivation and gives rise to certain normative dispositions (Oakley & Cocking, 2001, p. 25). For example, I can internalise a regulative ideal of what it means to be a good student and thus be guided by that conception as I study insofar as it influences – or disposes – me to act in certain ways with regard to my professors, courses and assignments.

Regulative ideals can be general or specific. I may internalize a conception of what the golden rule means – to do to others as I would have them do to me. Or I may internalize a more specific conception of what it means to be a good teacher. This would involve developing a conception of the appropriate ends of education and being disposed to interact with students in ways consistent with those ends.

So far all the examples of regulative ideals I have provided have involved conceptions that persons might consciously consider when acting in relevant circumstances. When teaching I might consciously consider how I should respond to a student so as to be a good teacher. But Oakley and Cocking explain that we need not always be so aware of regulative ideals. To borrow an example from them, "I can be guided by [a] conception of jazz excellence when I am ensconced in playing jazz piano, without consciously formulating that conception as I play" (Oakley & Cocking, 2001, p. 26). A regulative ideal, then, is a conception of what is correct or excellent about a general or specific domain; a conception which disposes one to act in certain ways and which can be either consciously or unconsciously present.

Oakley and Cocking assess professional conduct in terms of its contribution to human flourishing. Put simply, good conduct is conduct consistent with the regulative ideals shaping a good role; a good role is one which contributes to the ends and functions of a good profession; and a good profession is one which is, "committed to a key human good, a good which plays a crucial role in enabling us to live a humanly flourishing life" (Oakley & Cocking, 2001, p. 74). At each "level", what counts as good – be it an action, role, profession or goal – is assessed in terms of its contribution to human flourishing. To borrow an example from Oakley and Cocking,

> ...if (as many suggest) it is appropriate to take serving health as the central goal of medicine, then given the importance of health for human flourishing, medicine would clearly count as a good profession on this approach. And, for example, given a general practitioner's concern with the broad health needs of their patients, the general practitioner's role within medicine would seem to count as a good professional role. (Oakley & Cocking, 2001, p. 74)

On Oakley and Cocking's view then, professionals can contribute to human flourishing – and thus conduct themselves well – by contributing to professions which contribute to human flourishing. And good professional roles are those guided by regulative ideals disposing professionals to contribute to human flourishing in that way. Good professional conduct and good professional roles must be understood in the context of the goals of the professions within which they are situated and with consideration to how those goals enable persons to live fully flourishing lives.

Oakley and Cocking suggest that their virtue theory approach fits especially well with the assessment of professional roles since, in contrast to mere occupations, practitioners in professions deal with goods that play a “crucial *strategic* role in our living a flourishing life for a human being” (Oakley & Cocking, 2001, p. 79). But given the importance of institutions such as governments, banks, religious institutions and other institutions for our lives, I think their approach can be generalized to institutional contexts and roles more broadly, whether professional in nature or not.

Section 4.3: Integrated Virtuous Character in Institutional Contexts

Integrated, virtuous character is best understood as a complete system of institutional functions, roles and policies which dispose an institution and its role-holders to contribute to human flourishing. I will argue to this conclusion by extending Oakley and Cocking’s view of virtuous professional roles and Potter’s requirements for individual trustworthiness to institutional contexts.

Given Oakley and Cocking’s view of virtuous professional roles, virtuous institutional roles can be understood as those which dispose role-holders to aid in the fulfilment of good institutional functions. And good institutional functions will be those which contribute to human flourishing when they are fulfilled well. For example, consider the functions of religious institutions. I take it that one of the functions of such institutions is to aid in the spiritual development and wellbeing of persons. Given that spiritual development and wellbeing are important to living flourishing lives, religious institutions serve a good function insofar as they aid spiritual development and wellbeing well. Further, roles in religious institutions can contribute to human flourishing and so be virtuous insofar as they aid in the fulfilment of that spiritual development and wellbeing. In section 4.1 I said that in addition to functions and roles, policies affected the way institutions fulfilled their functions. So in addition to virtuous institutions involving roles which contribute to their good functions being fulfilled well, they will involve policies which do the same.

What it means for institutional roles and policies to dispose role-holders and institutions to perform good functions well requires explanation. I now turn to provide that explanation by applying the requirements for trustworthiness that I adapted from Potter in *Chapter 2* to institutional contexts. As I explained in C*hapter 2*, trustworthy individuals are disposed to giving assurances of trustworthiness; taking epistemic responsibility seriously; being sensitive to the particularities of others; responding properly to broken trust; dealing with hurt in relations in ways that sustain connection with others; working to resist and reform non-virtuous institutions; responding to the disempowered and exploited in a way that does not further their disempowerment or exploitation; being committed to mutuality; and working to sustain connection in intimate relationships without endangering mutual flourishing. And they also have a number of "other regarding" virtues. Those requirements identify ways in which trustworthy individuals respond well to those counting on them. I think they can also be applied to institutional contexts to identify ways that institutions can fulfil good functions well. That is, trustworthy institutions will not only be disposed to fulfil good functions; they will be disposed to do so in good ways.

First, institutions can be disposed to *give assurances of their trustworthiness* by establishing procedures for publicising information about policies and features of roles which serve those counting on them. For example, if a hospital were to put in place a new and improved policy for reviewing and improving the care of its patients, inform the patients that the policy had been put in place, and then show that the hospital was in fact enacting the policy, it could indicate that it cares about patient well-being. But in doing so it would also indicate that it cares about informing patients of steps it is taking to be a better hospital.

Second, institutions can *take epistemic responsibility seriously* by working to counter prevailing assumptions – both about the institution and its role-holders and about outsiders – which are inconsistent with their functions. For example, in the child migrant case I presented in section 2.2, to be epistemically responsible those persons in institutions entrusted with the "care" of the migrants would need to have worked to counter assumptions that the children they are "caring" for are ill behaved "bad" children who either do not require or are unworthy of affection.

Third, institutions can be disposed to be *sensitive to the particularities of others* by having policies encouraging role-holders to consider the position of specific others with whom they interact. In the child migrant case, institutional workers could have been required to discover as much as was possible about the situation that brought each child into their care and to take account of significant aspects of the child's history that may affect how they should be cared for.

Fourth, an institution can be disposed to *respond properly to broken trust* by having procedures in place for taking responsibility for any damages done to trust by its role-holders or by the institution as a whole. For example, in the Catholic Church case when the abuse survivor approached the diocese about the abuse she had suffered, the offending priest was quietly removed from his parish and given the opportunity to retire; little was done within the church to accept responsibility for what had happened (Fleming, 2007, p. 160).[42] In contrast, if the institutional church had publicly acknowledged the wrong done and shown that they were taking steps to prevent it from happening in the future, the survivor's broken trust might have been repaired to some extent.

[42] When the survivor took the diocese to court for her abuse, they offered her a settlement which involved a "no-fault clause" (Fleming, 2007, p. 162). As the survivor writes, "They were agreeing to this settlement without accepting any fault for what happened. I could never sue them again, and I could never publicly disclose any information about this case. It was a gag order, and I knew it" (Fleming, 2007, p. 162).

Fifth, institutions can be disposed to *deal with hurt in relationships in ways that sustain connection with others* by implementing processes for giving uptake to those they hurt.[43] For example, if a patient in a hospital is hurt by the care given to her by hospital personnel, her trust in the hospital may not necessarily be broken. She might just think her being hurt was a mistake that is unlikely to happen in the future either to her or to other patients. Still, the hospital could acknowledge the hurt done to the patient, take responsibility for it, and take steps to help the patient with any effects of what happened to her while at the hospital.

Sixth, institutions can *work to resist and reform non-virtuous institutions* they are a part of. Institutions can be internally differentiated and those institutions that are a part of larger institutions may work to resist and reform their non-virtuous "parent" institutions. For example, a committee for social inclusion in a university may work to reform their university's discrimination prevention policies in order to make the university a more inclusive "parent" institution. Whether the committee succeeds in reforming the university's policies or not, it would be working to reform a non-virtuous institution.

Seventh, institutions can be disposed to *respond to others in ways that do not exploit, or further exploit, them* by having policies which are sensitive to those who are already disempowered or exploited. For example, an institution could implement equal opportunity employment policies that seek to not further disadvantage any persons who are already disempowered.

Eighth, institutions can be disposed to encourage *mutuality* by establishing ways for individuals to contest power imbalances within the institution and between individuals and the institution. Institutions and their role-holders may be given special authority – e.g. judges' ability to sentence persons to prison – and so it seems the relationship between individuals

[43] This requirement is different from the fourth requirement – that the trustworthy respond properly to broken trust – because not all hurt results in broken trust. Rather, the trustworthy will respond well to hurt occurring in a relationship whether the trust therein is damaged or not.

and them is not meant to be mutual. But by establishing ways for individuals to contest power imbalances institutions can be disposed to conduct themselves well. For example, a court of law may implement appeals processes so that individuals can contest rulings passed by its judges. Special authority may be given to institutions and their role-holders but trustworthy institutions will take steps to guard against that authority being used wrongly.

Potter’s ninth requirement – that the trustworthy are disposed to *sustain connection in intimate relationships without endangering mutual flourishing* – does not translate straightforwardly to institutional contexts. The relationship between individuals and institutions is not characterised by intimacy. However, there are institutional analogues to Potter’s requirement. Institutions can establish and implement policies which encourage good working relationships amongst its role-holders. For example, by implementing an anti-bullying policy an institution could express a commitment to establish good workplace relationships that aid in the flourishing of its staff.

Tenth, institutions may also be disposed to fulfilling their functions well by *having a number of other virtues.* In the policies they create, institutions can express, among other things, a regard for the good of others and compassion for them. For example, a hospital may care for its patients compassionately insofar as it has policies enabling patients to defer payments for services. And the doctors and other role-holders within a hospital may express virtues in the way they treat their patients with respect, compassion and good care.

The above requirements identify ways in which institutions can be disposed to fulfil good functions well. By having the right roles, policies and processes in place, institutions can be disposed to contribute to human flourishing, not only by performing functions which contribute to human flourishing but by conducting themselves in ways that promote flourishing. Trustworthy institutions do not just fulfil good functions; they fulfil good functions *well*.

Given my extension of Oakley and Cocking's view and Potter's requirements to institutional contexts, there are at least two types of ways in which institutions can fail to be trustworthy. They can fail to have integrated, virtuous character if the roles and policies in an institution aid in the performance of functions that do not contribute to human flourishing. For example, consider the Nazi government of the 1940s. I assume that one of the ends or functions of that government was to ensure the ethnic purity of the Germanic race, a function they set about achieving through a process of genocide. And presumably, roles and policies in the Nazi government were constructed to aid in the fulfilment of that function. Insofar as Nazi roles and policies disposed those involved in the Nazi government to aid elimination of certain people groups, the Nazi government could be said to have integrated character. But the function of the Nazi government that I have identified does not contribute to human flourishing.[44] And so while that government's character may have been integrated, it was not virtuous and trustworthy.

Institutions can also fail to have integrated, virtuous character if they have good functions but lack roles and policies which aid in those functions being performed well. Their functions may dispose them to contribute to human flourishing, but the roles and policies disposing them to fulfil those good functions would be lacking. Some of what makes up their character – their function – would dispose them to contribute to human flourishing but their character would not be integrated; some parts would be disposing the institution to contribute to human flourishing while others – the roles and policies – would not. Rather, for institutions and role-holders to be trustworthy, they must have character that is, like trustworthy individuals, both integrated and virtuous. They must have a complete system of functions, roles, policies and processes which dispose them to contribute to human flourishing.

A trustworthy institution or role-holder may occasionally fail in contributing to

[44] A member of the Nazi government would have perhaps thought that ethnically cleansing the state would contribute to human flourishing. But such a view of human flourishing is deeply flawed.

human flourishing. But such lapses must be occasional and local, not systemic. For example, in the Catholic Church case, if the abusive priest’s conduct resulted merely from his individual character and was not enabled by his institutional role or the policies shaping his role, then the institution of the Catholic Church might be trustworthy. But his abuse of power was enabled by a systematic failing in the Catholic Church to guard against such abuse. And so his actions do not only express the failure of an individual to contribute to human flourishing but also the failure of an institution to do so. Further, if the Catholic Church had taken responsibility for the abuse that had occurred within the institution and taken steps to reform the relevant roles and policies to guard against future abuse in the church, it would have shown that it was concerned about becoming more trustworthy.

To recapitulate, trustworthiness is characterised by a kind of integrated, virtuous character disposing the trustworthy to act in ways which contribute to human flourishing. This is the case whether the trustworthy are individuals, institutions or role-holders. But what it means for individuals to have such trustworthy character and what it means for institutions and role-holders to have it are slightly different. Trustworthy individuals have character such that they are disposed to respond to those counting on them in ways that contribute to human flourishing across various situations. In institutional contexts, the trustworthy have functions, roles and policies which dispose them to contribute to human flourishing as they fulfil good functions well.

In this chapter I have shown that it is possible to give substantive accounts of the concepts of trust, trustability and trustworthiness in institutional contexts. The cluster concept of trust is applicable to interactions involving institutions and role-holders. It is possible for individuals to know enough to have good reason to trust institutions and role-holders. And institutions can have integrated, virtuous character. With the analysis of trust, trustability and trustworthiness I have provided across the last three chapters in hand, I now turn in C*hapter 4* to explicate the concept of betrayal.

Chapter 4: Betrayal

Section 1: Introduction

Philosophers commonly use the concept of betrayal to distinguish trust from mere reliance (Baier, 1986, p. 235; Hieronymi, 2008, p. 215; Holton, 1994, p. 66; McGeer, 2002, p. 33; McLeod, 2011; Walker, 2006, p. 85; Wright, 2009). On these views, trust can be betrayed while mere reliance cannot. This was the point of Annette Baier's famous example of Immanuel Kant's neighbours who rely on his daily walks to set their unreliable clocks (Baier, 1986, p. 235). "Kant's neighbors who counted on his regular habits as a clock for their own less automatically regular ones might be disappointed with him if he slept in one day, but not let down by him, let alone had their trust betrayed" (Baier, 1986, p. 235).

The above philosophers employing the concept of betrayal in their analysis of trust have not explained what exactly they take betrayal to be. But if the concept of betrayal is to aid in the philosophical analysis of other concepts – like trust and mere reliance – it needs to be defined and explained. As I pointed out in the introduction of this thesis, a clear account of betrayal is also needed to illuminate the damage that betrayal can inflict on trust. I discuss that damage in *Chapter 5*.

In this chapter I provide a preliminary analysis of betrayal and develop two main arguments: that betrayal is best understood as a type of disloyalty; and that not all betrayers are blameworthy. I begin by presenting three cases of betrayal in section 2. From those cases I identify the following preliminary features of betrayal: harm; deliberate use of a relationship for one's own gain; deception; and disappointment of expectations.

In section 3 I show that while all the features identified in section 2 can be involved in instances of betrayal, only disappointment of the normative, "constitutive" expectations that shape a particular relational domain is a necessary feature of betrayal. I distinguish such expectations from "predictive" expectations. Victims of betrayal may not have their predictions of an offender's behaviour disappointed, but normative expectations regarding what is appropriate behaviour for a relational domain will be disappointed in cases of betrayal. Harm, deliberate use of a relationship for one's own gain, and deception, may also be present in cases where a person betrays another – and deception may even aid a person in successfully betraying another – but they need not always be involved in instances of betrayal.

In section 4 I argue that given my preliminary analysis, betrayal is best understood as a type of disloyalty. I follow Simon Keller in understanding disloyalty as doing an affront to someone by failing to meet a normative expectation that you will take their side because you have a special relationship with them (Keller, 2007, p. 211). I show that the constitutive expectations disappointed in the cases of betrayal presented in section 2 are the kinds of expectations that are not upheld when one is disloyal. This might suggest that we should equate betrayal with disloyalty. But I show that it is possible to be merely disloyal without betraying. Rather, I explain betrayal as a specific type of disloyalty which does not just do an affront to someone you are in a special relationship with but which *undermines* that special relationship. While all disloyalty involves a failure to uphold normative expectations, betrayal involves failing to uphold normative expectations in a way that compromises a special relationship and can even prevent it from being sustained.

My first main argument – that betrayal is best understood as a type of disloyalty – leaves the moral status of betrayal open and does not answer whether or not those who betray are always blameworthy. But the apparent blameworthiness of betrayal needs to be addressed. In section 5 I focus on the morality of betrayal and argue that betrayers are not always blameworthy. I show that there are excusing conditions for betrayal.

Given the understanding of betrayal argued for in sections 2-5, I show in section 6 that both trusters and those merely relying on others can be vulnerable to betrayal. But while vulnerability to betrayal is not limited to trust I show that there is still reason to associate betrayal with trust. Trust, but not mere reliance, can establish the special relationships undermined by betrayal. In order for those merely relying to be vulnerable to betrayal, features beyond mere reliance are required.

Section 2: Betrayal Phenomena

I begin with three uncontroversial cases of betrayal in different relational domains: friendship, marriage-like relations[45] and citizen-nation relations. These domains mark areas of human life in which betrayal often occurs. Betrayal can occur elsewhere, but perhaps the most familiar and frequent cases arise between friends, partners and between citizens and nations. After introducing each case I identify central features of that case which I will use to establish a set of features that a satisfactory understanding of betrayal will need to account for.

Judas: Betrayed Friendship

The Book of Mathew and *The Book of Mark* in *The Bible* tell the story of Jesus being betrayed by a close friend named Judas (Mathew 26; Mark 14, Today's New Living Translation). Judas arranges to give up Jesus to religious leaders. "Now the betrayer had arranged a signal with them: 'The one I kiss is the man; arrest him and lead him away under guard'" (Mark 14:44).

[45] By "marriage-like relations" I mean intimate, exclusive partnerships between individuals. The second example I present in this section involves the betrayal of a marriage partner by one who is sexually unfaithful to his wife. The importance of that case does not stem specifically from the fact that those involved were legally married. Rather, it is that one partner was sexually unfaithful to the other. And it is possible for partners to be sexually unfaithful to each other whether they are legally married or not. They need only violate the norms governing an exclusive, committed relationship. In that exclusivity and commitment, such partnerships are similar to western, monogamous marriage. And so when referring to the domain that the second case I present occurs in, I will talk primarily of 'marriage-like relations.'

Judas' actions take place just hours after Jesus tells his close friends – including Judas – that one of them is going to betray him. While eating their last meal together, Jesus says, "...The one who has dipped his hand into the bowl with me will betray me" (Mathew 26:23). When Judas asks if he is the one Jesus is talking about, Jesus responds positively saying, "You have said so" (Mathew 26:25b). After Judas hands Jesus over to the religious leaders he is arrested and put to death (Mathew 26, 27; Mark 14, 15).

Jesus' prediction of his betrayal is significant. As I will explain in section 3, betrayal can involve disappointment of expectations. But Jesus' accurate prediction of his betrayal shows that it is possible to be betrayed without having at least what I will call personal, "predictive" expectations disappointed.

The *deliberate use of a relationship to further one's own interests* is a central feature of the Judas case. His close relationship with Jesus enables him to hand Jesus over to those who want to kill him. Further, we are told that Judas hands Jesus over for personal financial gain (Mark 14:11). Whether Judas entered into a relationship with Jesus for manipulative reasons or not, Judas deliberately uses his relationship with Jesus to further his own interests.

That Judas puts Jesus in a position of imminent physical *harm* as he hands Jesus over is also a central feature of the case. Judas does not just hand Jesus over to people that Jesus dislikes or avoids. He hands Jesus over to persons who want to kill him.

Deception is also an important feature of the Judas case. Judas gives Jesus up to the religious leaders with a deceptive kiss. Judas uses an expression of union, friendship and affection to close a deal to hand his friend over to his enemies. This deception also features in the next case of betrayal, involving infidelity.

Tiger Woods: Betrayal in Marriage

In November 2009, news broke that professional golfer Tiger Woods had been cheating on his wife, Elin Nordegren (McCahill, 2009). Woods has acknowledged his infidelity and its impact on his wife, family, athletic sponsors and fans. In a press conference apology he said,

> Now every one of you has good reason to be critical of me. I want to say to each of you, simply and directly, I am deeply sorry for my irresponsible and selfish behavior I engaged in. I know people want to find out how I could be so selfish and so foolish. People want to know how I could have done these things to my wife Elin and to my children. And while I have always tried to be a private person, there are some things I want to say. Elin and I have started the process of discussing the damage caused by my behavior. As Elin pointed out to me, my real apology to her will not come in the form of words; it will come from my behaviour over time. We have a lot to discuss; however, what we say to each other will remain between the two of us. I am also aware of the pain my behaviour has caused to those of you in this room. I have let you down, and I have let down my fans. For many of you, especially my friends, my behaviour has been a personal disappointment. To those of you who work for me, I have let you down personally and professionally. My behaviour has caused considerable worry to my business partners. [...] The issue involved here was my repeated irresponsible behaviour. I was unfaithful. I had affairs. I cheated. What I did is not acceptable, and I am the only person to blame. I stopped living by the core values that I was taught to believe in. I knew my actions were wrong, but I convinced myself that normal rules didn't apply. I never thought about who I was hurting. Instead, I thought only about myself. I ran straight through the boundaries that a married couple should live by. I thought I could get away with whatever I wanted to. I felt that I had worked hard my entire life and deserved to enjoy all the temptations around me. I felt I was entitled. Thanks to money and fame, I didn't have to go far to find them. I was wrong. I was foolish. I don't get to play by different rules. The same boundaries that apply to everyone apply to me. I brought this shame on myself. I hurt my wife, my kids, my mother, my wife's family, my friends, my foundation, and kids all around the world who admired me. [...] Parents used to point to me as a role model for their kids. I owe all those families a special apology. I want to say to them that I am truly sorry [...] Finally, there are many people in this room, and there are many people at home who believed in me. Today I want to ask for your help. I ask you to find room in your heart to one day believe in me again. Thank you. (Woods, 2010)

As Woods says, his actions have effects beyond his relationship with his wife. Other people’s dependence on Woods is damaged in complex ways. However, my main concern is the damage his betrayal has done to the marriage partnership between himself and Nordegren.

Woods’ *disappointment of expectations* is a central feature of this case. Whether or not Nordegren suspected Woods’ infidelity, and whether or not she grew to expect him to be unfaithful, at some level Woods’ actions disappoint what is to be expected of a husband. In section 3 I will explicate more fully the expectations that are disappointed in cases of betrayal.

At some level there is also a significant element of Woods' infidelity *harming* Nordegren. The harm inflicted may not be physical like that done to Jesus in the Judas case. But not all harms are physical. Woods' actions arguably harmed Nordegren's emotional and social wellbeing. She may need to work through feelings of abandonment or insecurity. And she must face leaving her home with the general public knowing what her husband has done. While most who hear her story may empathize with her, she must still live with people talking about her personal business.

Also, like the Judas case, Woods used *deception* in his relationship with Nordegren. While the particulars about how he carried out his infidelity are undisclosed, it is highly likely that Woods would have had to lie to maintain his relationships with both Nordegren and others.

Disappointment of expectations, harm and deception are not limited to betrayal between individuals. In the next case, I show that betrayal involving similar features can occur between individuals and institutions.

The Cambridge Spies: Betrayal in Citizen-Nation Relations

In the 1930s young men were recruited from Cambridge University to be Soviet spies ("The Cambridge Spy Ring," 1999). Informally led by Harold "Kim" Philby, a group of five men, known as the Cambridge Spies moved into jobs in British Intelligence. Along with Philby, Guy Burgess, Donald Maclean, Anthony Blunt and John Cairncross passed valuable, and sensitive, information to the KGB. Of the information the spies passed, most notable is information leaked by Cairncross that enabled Soviet spies to change their codes just as British Intelligence were about to crack them ("The Cambridge Spy Ring," 1999). The work of the Cambridge spies was significant. It is thought that the information they leaked became the foundation of the Soviet nuclear programme ("The Cambridge Spy Ring," 1999).

As in the Judas and Woods cases, *deception* and *harm* are evident in the Cambridge Spies case. To reach their goals the spies would have had to exercise deception. And in sharing information with the Soviets, they were putting the nation they were supposed to protect, and their fellow British citizens, in a position where they could be harmed.

Also like Judas, the Cambridge Spies *deliberately use a relationship to further their own interests*. They manipulate their access to sensitive British intelligence for their own political, and probably monetary, ends.

I think it is also likely that in committing treason the spies *disappoint expectations* held by their fellow citizens and by other British Intelligence personnel. While it is unlikely that most citizens would actively consider what they expect of intelligence agents, it is plausible that citizens would have a background expectation that such agents would not divulge sensitive information to other countries. And I think it is quite likely that other British Intelligence personnel would have expected the defecting agents not to have acted as they did.

While the above cases of betrayal occur in different domains of relationship, they involve some similar central features. Harm or potential harm; deliberate use of a relationship for one's own gain; deception; and disappointing some type of expectations, are recurring features in the cases. A satisfactory understanding of betrayal will need to account for these features.

Section 3: A Preliminary Analysis of Betrayal

Not all the features identified above are necessary for betrayal. It is possible to betray without harming another, without deliberately using a relationship for one's own gain, and without being deceptive. However, it is not possible to betray without disappointing some expectations. As I explain, it is not mere disappointment of expectations in general that is necessary for betrayal, but disappointment of constitutive expectations shaping relational domains.

Section 3.1: Harm

While harming someone directly or putting them in a position where they are likely to be harmed are common features of the cases I presented in section 2, they are not distinctive of betrayal. It is possible to betray without harming someone and it is possible to harm someone without betraying them. To see this, consider a situation where I share a piece of information with a close friend. The information is not especially important but it is something I am embarrassed about and I ask her to keep it secret – perhaps it is that my favourite wine comes in a box! Then, when I am at a party with my friend, under no pressure from others, she embarrasses me by telling everyone about my secret enjoyment of cheap wine. I think it would be too strong a claim to say that my friend harms me or puts me in a position of harm by telling my secret to others, but I do think that by telling my secret she betrays me. So while harm may be a feature of betrayal, it is not always present.

Harming someone is also not enough to constitute an act of betrayal. It is possible to harm someone, or put them in a position where they are likely to be harmed, without betraying them. For example, if a careless driver strikes a pedestrian, the driver harms the pedestrian but I do not think it would be correct to say that he *betrays* her. It may be thought that the harm of hitting a pedestrian is somehow different from the harms involved in the cases of betrayal I introduced in section 2. Perhaps there is a certain kind of harm that *is* distinctive of betrayal. But I do not think that is correct. To show that the harm I identified in section 2 is not distinctive of betrayal, consider an alternate Cambridge Spies case. In this scenario the persons divulging British secrets are freelance hackers operating outside of the UK and selling their information to the Soviets. Such hackers would still be putting Britain and her people in a position of likely harm but I do not think it would be right to say that they were betraying Britain or her people. In section four, my account of betrayal as a type of disloyalty explains why the freelance hackers' actions do not constitute betrayal: their actions do not undermine a special relationship between them and Britain or her people.[46]

It may be thought that the reason the freelance spies harm Britain and her people without betraying them is that Britain and her people are not trusting the freelance spies for anything. They are presumably not even aware of them. In contrast, in the original Cambridge Spies case there is reason to understand the relationship between the spies and those they betray as one of trust. The spies are trusted with sensitive British intelligence. So perhaps the reason that the Cambridge Spies betray Britain and her people while the freelance spies do not is because harming a *truster* is distinctive of betrayal but merely harming another non-truster is not. But I do not think that causing harm to a truster is distinctive of betrayal either. Such a view fails to account for the difference between mere let down and betrayal. It is possible for a trustee to let me down and as a result harm me, or put me in a position of immanent harm, without betraying me. For example, I trust my flatmate to turn off the stove when he is done cooking. If he fails to turn the stove off and our apartment building burns to the ground, he will have let down my trust and harmed me, or at least my possessions, but I do not think it would be right to say that he has betrayed me.

Rather than betrayal being characterised by harm, perhaps it is primarily a matter of using a relationship with another to further one's own interests. Such an understanding of

[46] In section 4 I argue that betrayal is best understood as a failure to uphold normative expectations in a way that undermines special relations. It may be thought that by undermining a special relationship with another, betrayers harm those they were in relationship with. And indeed, having one's special relationship with another undermined by them may be experienced as harm. For example, while Wood's infidelity did not harm Nordegren physically, she may feel that he has harmed her insofar as he has severely damaged their relationship. If, as I argue in section 4, betrayal involves disappointing normative expectations in a way that undermines special relations, and if that undermining can be experienced as harm, then it may seem that betrayal does in fact involve harm. But, while betrayal can involve harm – and while the undermining of a relationship can be experienced as harm – I still do not think betrayal should be understood primarily in terms of harm. Even if it turns out that all victims of betrayal experience the undermining of their relationship with another as a kind of being harmed, I do not think harm primarily characterises betrayal. Rather, as I explain in section 3.3, the disappointment of normative expectations is the distinctive feature of betrayal.

betrayal could account for the feature of harm that is often involved in betrayal but not distinctive of it. Betrayers might harm others sometimes because they don't really care about them as much as they care about reaching their own goals. But as I will show in the next part of this section, using one's relationship with another to further one's own interests is only a contingent feature of betrayal.

Section 3.2: Deliberate Use of a Relationship to Further One's Own Interests

Both the Judas case and the Cambridge Spies case involve persons deliberately using relationships for their own gain. Judas uses his friendship with Jesus for profit. And the spies presumably use their relationship with the British government – their roles as intelligence officers – for their own political interests or perhaps because of financial gain. Rodger L. Jackson has argued that such deliberate, or intentional,[47] use of a relationship for one's own gain is a necessary feature of betrayal. As he puts it, betrayal involves taking an "instrumental view" of a relationship with another (Jackson, 2000, p. 85). He also claims that deception is a necessary feature of betrayal. On his view, betrayers must use deception to successfully carry out the actions stemming from their instrumental attitude to a relationship (Jackson, 2000, p. 85). I will show that Jackson's view is unsatisfactory for two reasons: it overstates the role of deception in betrayal – it is possible to betray without necessarily being deceptive; and it cannot account for cases of betrayal where an instrumental view is not taken.

Apart from saying that betrayal is an, "...intelligible and purposive event" (Jackson, 2000, p. 84) and that it, "...does not just happen accidentally" (Jackson, 2000, p. 85), Jackson does not explain what he takes intentional, or "deliberate", action to be. But getting a clear understanding of such action is important for understanding various types of betrayal. As I

[47] Jackson uses "deliberate" and "intentional" interchangeably. An example is when he analyses Willoughby's violation of his relationship with Marianne in Jane Austen's *Sense and Sensibility*. "The calculated and deliberate campaign that Elinor and Marianne *thought* Willoughby was engaged in has the intentional quality we associate with betrayal" (Jackson, 2000, p. 85). I will follow Jackson in talking about 'deliberate' action and 'intentional' action interchangeably.

will argue below, contrary to Jackson's view, not all betrayal is done intentionally.

For an explanation of intentional action I look to G.E.M. Anscombe. She understands intentional action to be such that the question of "why?" is appropriate (as cited in Stout, 2005, p. 20). An intentional action may be done without an agent thinking she is doing it for any particular reason, but the behaviour is such that it is appropriate to inquire of the agent as to why she behaved as she did. When asked why she behaved the way she did, the agent may say, "no reason really" but the behaviour is such that it is at least appropriate to ask her the "why?" question.

According to Jackson betrayal is not just intentional, it also involves taking an instrumental view of a relationship one is in. Betrayers take a relationship to be a means to their own goals, and they see the other person and the relationship as a thing to be manipulated.

> A betrayer sees the relationship of trust in fundamentally instrumental terms. The relationship is the medium through which a betrayer creates an effect or obtains a prize. The betrayer is engaged in the relationship; its healthy existence is vital to the successful completion of the intended effect. A betrayer cannot stand aloof; like a craftsman, he must know and manipulate his medium. (Jackson, 2000, p. 85)

Jackson supports his view with an example from Jane Austen's *Sense and Sensibility*. He explains the actions taken by Willoughby in his relationship with Marianne as abandonment rather than betrayal.

> Willoughby does not see his trust relationship with Marianne in instrumental terms; his actions are not the calculated maneuvering of a traitor setting up his victim. What Willoughby does instead is to unilaterally withdraw from the relationship, indifferent to (or at least insufficiently concerned with) the effect on Marianne. He *abandons* his care of the relationship of trust. While this is clearly a violation of her trust, it is not betrayal. He is not engaged in the relationship to manipulate it; rather, he disengages from it entirely. (Jackson, 2000, p. 86. Emphasis is Jackson's.)

On Jackson's view, then, betrayal is an intentional manipulation of a relationship that one takes an instrumental stance toward. Willoughby does not see his relationship with Marianne as something to be manipulated and so, on Jackson's account, does not betray her. I disagree with Jackson and think Willoughby's actions do in fact amount to a betrayal of Marianne. After explicating my view of betrayal as a type of disloyalty in section 4, I explain why Willoughby's actions constitute betrayal.

Jackson's view accounts for cases where persons enter relationships deliberately for the purpose of betrayal. The treachery done by the Cambridge Spies is such a case. After they were recruited from Cambridge by the Soviets, the spies entered their roles in British intelligence for the purpose of passing sensitive information from England to Russia.

Jackson's view also accounts for cases where persons come to take an instrumental view of a relationship after it has begun. For example, Judas did not enter his relationship with Jesus for the express purpose of handing him over to the religious leaders. Rather, in the *Book of Mark* we are told that Judas enters into close relationship with him because Jesus appoints Judas to be one of twelve "apostles" who would be closest to him (Mark 3:13-19). If Judas sees his relationship with Jesus as a thing to be manipulated, he only comes to that perspective later. So while both the Cambridge Spies and Judas use the relationships they are in to further their own goals, the spies enter into their relationship for that purpose while Judas changes to take that stance once he is already in the relationship with Jesus.

On Jackson's view, betrayers employ deception in their manipulation, regardless of whether they enter a relationship for the direct purpose of manipulating it or not. As such, Jackson infers that deception is a necessary part of betrayal.

> ...in order to achieve the goal, betrayers must lie or mislead the truster about their intentions at critical moments in the relationship. This is so because their objective is not the care of the object of trust, whatever it is, but the use of the relationship to achieve a goal extrinsic to it. (Jackson, 2000, p. 85)

I agree with Jackson that deception can be a practical tool aiding the betrayal of another. A case in point here is Judas' deceptive kiss. But it is possible to betray without using deception. Imagine an alternate Cambridge Spies case in which the spies do not try to hide what they are doing. They consider themselves to be heroic captains of treason; willing to go down with their sinking nation as they leak information to the Soviets. They broadcast the sensitive information on international radio. They betray their homeland but without hiding; without deception. When confronted they say, “Yes, we have betrayed our country to the Soviets.” I think the actions of such spies should be understood as instances of betrayal even though they do not use deception.

Betrayal might involve deception of persons other than its direct victims. For example, recall the survivor of clergy abuse I discussed in *Chapter 3*. The secrecy that existed within the Catholic Church about that abuse was a type of deception even if the survivor was not deceived by the priest. At the start of her relationship with the priest, the survivor was unaware of its abusive nature, but presumably at some point she became aware of what was being done to her. But whether the priest deceived the survivor or not, he deceived the rest of his parishioners as well as other clergy in the Catholic Church. So it may be thought that betrayal always involves some type of deception – either of a victim or of third-parties. But I think the heroic spies case I presented above is also a counterexample to that claim. The spies blatantly announced that they were betraying their nation. It is plausible that they might have betrayed their nation to the Soviets despite not deceiving their victims – e.g. Britain and her people – or third-parties such as other countries or The United Nations. Success would be more difficult, but it is at least theoretically possible for them to betray the nation without being deceptive. They may just give Britain’s secrets to the Soviets while everyone watches. Because it is possible to betray without deceiving, deception is not necessary for betrayal.

Taking an instrumental view of a relationship is also unnecessary for betrayal. While the Cambridge Spies, Judas and the priest in the Catholic Church example may view their respective relationships instrumentally, not all betrayers do. Woods did not seem to see his relationship with Nordegren as something to be manipulated. Rather, he just doesn't seem to notice, or take account of, his relationship with her at all.

> I never thought about who I was hurting. Instead, I thought only about myself. I ran straight through the boundaries that a married couple should live by. I thought I could get away with whatever I wanted to. I felt that I had worked hard my entire life and deserved to enjoy all the temptations around me. (Woods, 2010)

While Woods does not take an instrumental view of his relationship with Nordegren, I still think it is appropriate to understand his infidelity as constituting betrayal.

The Woods case also calls into question Jackson's claim that betrayal is done deliberately. If we intercepted Woods on his way to meeting another woman and said to him, "You would betray your wife?" I do not think he would say, "Yes, that is what I mean to do." Instead, based on his claim that he didn't think about whom he was hurting, I think Woods might respond with something like, "I guess I hadn't really thought of it like that. I had pushed such considerations out of my head long ago. I was so caught up in my own desires that I hadn't even thought of her or our relationship." My claim is not that Woods acted unintentionally all together. On the contrary, I think he did act intentionally in pursuing partners other than his wife; but he did not go out meaning to betray. Applying Anscombe's understanding of intentional action to the case, the "why?" question is applicable to Woods but his answer to that question would not be, "to betray my wife." He acted deliberately but he did not *deliberately betray*. Rather, he let himself get into a situation that he was obligated not to get into and negligently betrayed his marriage partner. I take the Woods case to show that in addition to betrayal being done deliberately at times, it can also be done *negligently*. I will understand negligent betrayal as betrayal that is not done deliberately but that still violates an obligation. For example, consider a fictional variant on the Woods case. Imagine that Woods confides in a friend about his infidelity and asks the friend not to tell Nordegren about what he has been doing. The friend agrees to keep Woods' secret but one day, without meaning to, says something to Nordegren that allows her to deduce Woods' secret. The friend has betrayed Woods to Nordegren and has violated an obligation he had to Woods. But the friend does not deliberately mean to betray Woods. Rather, I think his betrayal of Woods is best described as negligent. The friend was obligated to keep Woods' secret but, without deliberately meaning to, neglected to do so.

My analysis may seem to let Woods and his friend "off the hook," but, as I will explain in section 5, negligent betrayal can be just as blameworthy as deliberate betrayal. For now I will only add that, although negligent, Woods' violation amounted to an extremely damaging betrayal. But his fault was not that he *deliberately* betrayed his wife but that he was not *more* deliberate in guarding against temptations to infidelity. He did not think about who he was hurting and he should have.

On my analysis, then, some betrayers may take an instrumental view of their relationships with others. But taking such a view is unnecessary for betrayal. And while some persons will betray deliberately, some may do so negligently. I now turn to a feature that *is* necessary for betrayal.

Section 3.3: Disappointment of Expectations

Disappointment of expectations is evident in the Woods case and the Cambridge Spies case. It is plausible that in both those cases the victims of betrayal would have been surprised when they found out what their respective betrayers had done. This suggests that disappointment of expectations is a necessary feature of betrayal. But the Judas case is a counter-example to that claim. Jesus predicts that Judas will betray him. And so in handing Jesus over to the religious authorities, Judas does not disappoint Jesus' expectations, and yet Judas' actions constitute a betrayal.

It might be thought that the Judas case is exceptional. Persons are not usually able to make predictions like Jesus did. It may turn out that, the Judas case aside, disappointment of expectations is a necessary feature of betrayal. But alternative counter-examples can be found. The relationship between English footballer Ashley Cole and his wife, singer Cheryl Cole, provides another example of betrayal where expectations do not seem to be disappointed.[48] Like Woods, Ashley cheated on his wife Cheryl. In February 2010 it was made public that Cole had cheated repeatedly on his wife even after it was known that he had done so in 2008 (Topping, 2010). In September 2010 the Coles divorced, but were they to stay together, Cheryl might predict that Ashley would cheat on her again (Boshoff, 2010). And if he did cheat on her again, Ashley would not be disappointing Cheryl's personal expectations of him; but he would still be betraying her.

Given the Judas case and the Cole case, it seems that disappointment of expectations is not necessary for betrayal. But different types of expectations are involved in relationships. There are personal predictions that we can make about how another will actually act. It is these predictive expectations that are disappointed in the Woods and Cambridge Spies cases but not in the Judas and Cole cases. There are also normative expectations, which I will call "constitutive expectations", which shape the domain of relationship in which interactions between the parties occur. These are expectations for which individuals can be held to account, given the domain of their relationship. As I understand them, constitutive expectations assume a context of moral responsibility. They involve an assumption that individuals in a relational domain are capable of knowing and responding to each other in ways that are specific and appropriate to the relational domain they interact in. Consider the domains relevant to the cases in section 2: friendship, marriage and citizen-nation relations.

I take the domain of friendship to involve norms of caring for another for her own sake, shared intimacy and desire for participation in joint pursuits.[49] Friends are expected to

[48] My thanks to Karina Jacoby for this example.
[49] I follow Dean Cocking and Jeannette Kennett's understanding of friendship as involving mutual affection, well-wishing, desire for shared experiences and intimacy (Cocking & Kennett, 1998, p. 26). I use "care for the

stick by each other because of their special relationship and can be held accountable for failing to do so; they value each other and their relationship directly. Friends bond and establish trust through being open and responsive to each other and as such are shaped by each other. And friends, either as a result of the previous two features or as an impetus for them, partake in joint pursuits, or at least desire to do so. They often fight for the same causes, work alongside each other and share the same interests, joys, fears and worries.

The constitutive expectations of Western, monogamous, marriage-like relationships involved in the Woods case are an extension of those shaping the domain of friendship. Such partnership involves caring for each other and for a shared relationship directly. The care involved in marriage-like relations is “direct” insofar as partners are committed to caring for each other and for their relationship as ends in themselves rather than as means to some other end. The expectation is not just that partners will care for each other or for their relationship when it is easy or desirable to do so. Rather, they are expected to care for each other even when doing so is hard or when the other does not deserve it.

Like close friendship, marriage-like relations also involve mutual self-disclosure. Partners are expected to be transparent with each other. This transparency is expressed in the sexual intimacy between partners as they allow themselves to be physically vulnerable to each other. It is also present insofar as partners are expected to not keep secrets from each other and to share belongings.

other for her own sake” to denote elements of mutual affection and well-wishing. Further, I follow Cocking and Kennett’s ‘Mutual Drawing’ account of intimacy in close friendship. Understood as mutual drawing, intimacy between close friends amounts to a shared openness to being affected by each other and by the relationship. This contrasts with the view that friendship involves a constitutive element of equal self-disclosure. (For a view of friendship as involving equal self-disclosure see Laurence Thomas’ “Friendship” (Thomas, 1987). Kennett and Cocking argue that understanding intimacy as equal self-disclosure misunderstands the nature of the self (Cocking & Kennett, 1998, p. 505). Close friends do not reveal who they are to each other but rather open themselves to having who they are be shaped by each other. They write, “It is not that I must reveal myself to, or see myself in, the other, to any great extent, but that, in friendship, I am distinctively receptive both to the other’s interests and to their way of seeing me” (Cocking & Kennett, 1998, p. 505).

Perhaps most relevant to the Woods case is the constitutive expectation that marriage-like partnerships are exclusive relationships. While one may have multiple close friends, in the Western context there is a normative expectation that partners relate to each other in exclusive ways. They mutually self-disclose, and are intimate with each other in ways that they are not with others.

The two main constitutive expectations I take to be shaping citizen-nation relations are that citizens will uphold the reasonable laws of their nation and that nations will uphold minimal obligations to their citizens. Whether it is thought that a group of citizens *will* actually uphold the laws of their nation, there is a normative expectation that they should do so unless they have good reason for thinking the laws unreasonable. Likewise, a national government may be corrupt and so not predictably expected to uphold its obligations to its citizens. But there is still the normative expectation that nations should fulfil at least minimal obligations to their citizens. My argument here does not depend on determining what will count as “reasonable laws” that citizens are expected to uphold. Nor does it depend on identifying what a nation’s “minimal obligations” to its citizens actually are. Rather, the claim is that some such expectations shape the citizen-nation relation. For my purposes I will take reasonable laws to be those which do not impinge on the human rights of citizens. And I will understand a nation to be minimally obligated to protecting its citizen’s human rights.

The above constitutive expectations in the citizen-nation relationship can also apply to persons that are neither full citizens nor non-citizens. For example, I am an international PhD candidate from the U.S. studying in Australia. I am not an Australian citizen and so I am not subject to the same expectations to which the Cambridge Spies would have been subject as British citizens. But I am subject to more expectations than the off-site hackers in my alternate spy case. If I were to somehow divulge Australian intelligence secrets to other countries I would be disappointing constitutive expectations of my relationship to the nation of Australia and her citizens. Even as an international student, I have placed myself in a special relationship with Australia for the duration of my study. This is evidenced in the requirement of my signing a declaration of compliance with Australian Values to receive my Australian Visa.

While not all the cases of betrayal I have discussed involve disappointment of predictive expectations, they each involve disappointment of some constitutive expectations. In selling Jesus and putting him in a position of harm, Judas fails to care for Jesus as a friend. In his deception, Judas does not share intimacy with Jesus; the little intimacy expressed with his kiss is untrue. In pursuing and sleeping with other individuals, Woods violates the constitutive expectation of marital exclusivity. Also, assuming that he was not forthright with Nordegren during the period of his infidelity, he fails to uphold expectations regarding mutual self-disclosure; he fails to be transparent with his wife. And in divulging sensitive information, the Cambridge Spies failed to uphold reasonable laws of their nation and breached norms governing their roles as British Intelligence personnel.

It is possible to betray without disappointing *all* of the constitutive expectations shaping a domain. A friend may betray another while still desiring to take part in joint pursuits with him. For example, Judas might have desired to live closely with Jesus while finding that desire overpowered by his love of money. Or an especially uncaring marriage partner might be unfaithful to her spouse while making no attempt to conceal her infidelity. Rather, she may be completely transparent in her betrayal. Such a marriage partner would be disappointing the constitutive expectation of exclusivity but not disappointing that of mutual self-disclosure and transparency. While disappointment of *all* constitutive expectations is unnecessary for betrayal, given the cases I provided in section 2, disappointment of at least *some* such expectations is necessary.

In this section I have shown that while harm, deliberate use of a relationship for one’s own gain, deception and disappointment of predictive expectations can be involved in instances of betrayal, they are not necessary for it. Instead, I have argued that disappointment of constitutive expectations is a necessary part of betrayal. This preliminary analysis makes a start at defining the concept of betrayal: it at least involves a disappointment of constitutive expectations. It remains to be explained how, in light of this, the concept of betrayal should be understood.

Keller’s explanation of loyalty and disloyalty provides insight into the importance of the disappointment of constitutive expectations to betrayal. The central feature in his account is a special relationship occurring between a person and the object of their loyalty. On his view, loyalty is primarily being motivated to “take the side” of something or someone because of that special relationship (Keller, 2007, p. 21). Conversely, disloyalty is failing to uphold a normative expectation that is present because of the special relationship which obliges you to “take [the other’s] side” (Keller, 2007, p. 211). In section 4 I show that there is good reason to understand the disappointment of constitutive expectations involved in betrayal as the same kind of failure that is involved in disloyalty. The constitutive expectations in relational domains identify a special relationship between those involved and oblige them to “take each other’s side.” As I will show, the concept of betrayal is therefore best understood as a type of disloyalty that undermines one’s special relationship with another.

Section 4: Understanding Betrayal as a Type of Disloyalty

On Keller’s view loyalty is taking something or someone’s side because one is motivated in a certain way to do so (Keller, 2007, pp. 15-20). By “taking something’s side” Keller means a kind of favouring something or having positive regard toward it (Keller, 2007, p. 21). “There are many ways in which you might take something’s side; taking X’s side may mean identifying with X, or treating X with concern, respect, veneration, reverence or love, or having any of a number of other attitudes of positive regard toward X" (Keller, 2007, p. 21). The focus of Keller's account is on the *motives* involved in loyally taking the side of someone or something.

Keller identifies three features of the motivation central to loyalty: emotional pull; response to a particular thing itself; and response to a "special relationship" with the thing one is being loyal to (Keller, 2007, pp. 15-16).

> Loyalty is the attitude and associated pattern of conduct that is constituted by an individual's taking something's side, and doing so with a certain sort of motive: namely, a motive that is partly emotional in nature, involves a response to the thing itself, and makes essential reference to a special relationship that the individual takes to exist between herself and the thing to which she is loyal. To be loyal to something is to have loyalty toward it. To act out of loyalty to something is to be driven to action by the motive just described. (Keller, 2007, p. 21)

The emotional pull and response to a particular thing are based on the special relationship between the loyal person and the object of his loyalty. We have an emotional pull toward those with whom we take ourselves to have a special relationship. And our understanding of the object of our loyalty is shaped by our special relationship with it. This special relationship at the centre of Keller's view is essentially a distinct connection that one takes oneself to have with the object of one's loyalty. For example, if I am a loyal, patriotic American citizen I will most likely feel an emotional pull toward the U.S. as a specific country; not just because I value the basic tenets of its constitution or appreciate American culture but because it is *my* country. Or to borrow an example from Keller, if I keep a promise out of loyalty to someone named "Bob" I do not just keep the promise because of the principle that promises should be kept – I might hold that principle as well but that is not what it means to keep the promise out of loyalty. Rather, it is to be motivated, "...as a person who made a promise to Bob" (Keller, 2007, p. 18). On Keller's view then, loyalty is primarily taking the side of something because of a special relationship one has to it. This special relationship is also central to his view of disloyalty.

Disloyalty is not the opposite of loyalty (Keller, 2007). Although loyalty is a matter of appropriate motives given a special relationship, disloyalty is not a matter of motives. Failing to be motivated to take the side of someone because of a special relationship is not necessary for disloyalty. For example, I may keep a promise to someone because of principles I value rather than out of loyalty to him. I may not be motivated by a special relationship to the other but rather just think it is always good to keep one's promises. On Keller's view I would not be being loyal. But I think it would be strange to call me disloyal. Rather, in the example, I am just not motivated in the way characteristic of loyalty. I am "non-loyal" but I am not disloyal.

Keller explains disloyalty in contrast to mere non-loyalty as a violation of an expectation that one will take the side of something with which one has a special relationship (Keller, 2007, p. 211). As he writes,

> To be disloyal to something – call it X – is to do X an affront by failing to meet a certain sort of expectation to which you are subject: namely, an expectation that exists by virtue of some special relationship between you and X, and that demands that you, in some sense or other, take X's side. (Keller, 2007, p. 211)

The expectation Keller identifies as being unmet in instances of disloyalty is not a mere predictive expectation but rather a normative one (Keller, 2007, p. 209). This is a similar distinction to the one I made in section 3 between predictive and constitutive expectations. Rather than disloyalty involving an affront to someone because you fail to do what she thinks you actually will do, it involves a failure to do what you ought to do given the special relationship you have with her. As Keller explains,

> There are two different ways in which we can speak of expectations. I can tell my class that I expect that everybody's paper will be handed in on the due date, and then tell my colleague that I certainly do not expect that everybody's paper will be handed in on time. In the one case, I am speaking of norms for behavior, or normative expectation; I am telling my students that it ought to be the case according to relevant standards, that everybody hands in their papers when they are due. In the other case, I am speaking of my predictions about what actually will happen, doubting that the standard referred to in the first case will be met. (Keller, 2007, p. 209)

Disloyalty is not just failing to act out of certain motives grounded in a special relationship but rather failing to take something's side *when you ought to do so given your special relationship to it.* For example, in some American towns football is very popular and loyalty to the local team is considered important. In such towns there is a normative expectation that citizens will support the local football team. This is not to say that persons merely predict that everyone will support the team but rather that there is the expectation that everyone *ought* to do so. If I grow up in such a town but support a football team from a neighbouring town, I may be seen as being disloyal to my town. And that may be the case regardless of my motives. I may not mean anything against the town by supporting another team; I simply prefer the other team. But I have a special relationship with that town which establishes an expectation that I will take the side of the town's team. In the eyes of the rest of the town's inhabitants, I ought to support the local team. However, if I have only recently moved to the town and do not support the local team, I may be non-loyal but I am not disloyal because as a newcomer I do not yet have a special relationship with the town. The expectation to support the team has not yet been established. As the newcomer I am not going against any normative expectations; the expectations don't apply to me yet.

Because disloyalty is not primarily a matter of motives but of normative expectations, it need not be done deliberately. Persons may purposely disappoint a normative expectation to "stick with" someone they are in special relationship with, but they need not. They can also do so negligently: out of distraction or forgetfulness. To borrow an example from Keller, consider a friend who expects you to fulfil the constitutive expectations of friendship. She assumes you will care for her, share some level of intimacy with her and desire to participate in joint pursuits. But for a time she does not meet any of those expectations herself. "She never has time to talk when it is you who needs to discuss things, she regularly stands you up, she is careless with your personal information, and so on" (Keller, 2007, p. 205). Such a friend is disloyal insofar as she fails to uphold normative expectations grounded in a special relationship which requires her to take your side. But she may not be meaning to do this. Rather, as Keller points out, “The reason why she is disloyal may be not that she means to be, but that she is preoccupied or distracted, or is going through a time during which she is forgetful and self-absorbed” (Keller, 2007, p. 205). Regardless of one’s motives then, I take disloyalty to be a failure to uphold a normative expectation obliging one to take something’s side because of a special relationship between oneself and that thing.

Keller’s understanding of disloyalty shares similarities with my preliminary analysis of betrayal. The disappointment of constitutive expectations that is involved in betrayal can be understood as a failure to uphold those normative expectations obliging one to take the side of something with which one has a special relationship. As I explained in section 3, the constitutive expectations I identified in the domains of friendship, marriage and citizen-nation relations are normative expectations. Friends may at times not actually think the other desires to take part in shared pursuits, but there is an expectation that friends ought to have such a desire, at least in some areas of life. Friends are not expected to want to do everything together but they are expected to desire at least some joint pursuits. Also, as I showed in the Cole case, marriage partners may not always predict that the other will remain faithful. Still, there remains a norm that marriage partners ought to be faithful to each other. And while citizens may doubt that their government will uphold minimal obligations to them, those obligations persist.

The normative, constitutive expectations involved in friendship, marriage and citizen-nation relations can be seen to exist in virtue of special relationships between the parties. They also oblige parties to those relationships to “side with” each other. A friend ought to care for another for her own sake, share intimacy with her and desire to participate in joint pursuits with her because they have a special relationship with each other. For example I am close friends with my flatmate Kenton. He may not always think I do the best job of caring about him but, given our friendship, there is a normative expectation that I ought to care for him. I ought to do that because Kenton is *my friend*; I have a special relationship with him. And because Kenton is my friend I am expected to side with him. That means doing the things that characterise the constitutive expectations shaping friendship: caring for him; sharing intimacy; and desiring to participate in joint pursuits with him.

The domains of marriage-like relations and citizen-nation relations can be understood in similar ways. Partners in marriage-like relationships are obliged to care for each other, take part in mutual self-disclosure and keep their relationship exclusive because they have a special relationship with each other. And because of that special relationship they are meant to side with each other. For example, Woods was expected to remain faithful to Nordegren not just because faithfulness is valuable or generally a good thing but because of their special relationship – because he is *her* husband and she is *his* wife. Because of that relationship they are expected to side with or remain faithful to each other. So too, citizens are expected to uphold the reasonable laws of *their* nation, and a nation is expected to uphold its obligations to *its* citizens, because of the special relation between them. In that domain, sticking with a nation or citizens can be understood just to mean upholding reasonable laws and obligations.

Given the above interpretation, the kind of expectations Keller has in mind were also relevant in the cases of betrayal I discussed in section 3. These cases involved normative expectations existing in virtue of special relationships obliging those involved to side with each other. And in betraying the respective parties to those relationships, Judas, Woods and the Cambridge Spies failed to uphold those expectations; they were disloyal.

The similarities between instances of disloyalty and betrayal might suggest that the concept of betrayal should be equated with disloyalty. But it would not be correct to understand betrayal in this way because it is possible to be merely disloyal without betraying.

Consider an example of disloyalty from Keller that is not betrayal: A mother and son are outside in winter and the mother becomes very cold. The son has the option to give her his jacket but refrains (Keller, 2007, p. 212). There is a special relationship between the mother and son that establishes, among other normative expectations, that they will care for each other. On Keller's view, the son's behaviour constitutes disloyalty. "[The son's] failure to take her side takes the character of a deliberate flouting of that expectation, and hence of an affront to [the] mother. It is disloyal" (Keller, 2007, p. 212). While the son in Keller's example is disloyal to his mother, I think it would be strange to say that in not giving her his jacket he *betrays* her.

Because it is possible to be merely disloyal without betraying, the concept of betrayal should not be equated with that of disloyalty. Instead, betrayal should be understood as a distinct *type* of disloyalty. Specifically, it is disloyalty that *undermines* special relationships. I will support this claim by way of answering the following question: why does the son in Keller's example not betray his mother while Judas, Woods and the Cambridge Spies do betray their respective victims? I think the difference is that the son's behaviour does not fail to uphold normative expectations in a way that undermines his special relationship with his mother while the behaviour of Judas, Woods and the Cambridge Spies does undermine the special relationships they are in.

By behaviour that "undermines" a special relationship I mean, behaviour that compromises the relationship and can even prevent that relationship from being sustained. I will explicate the notion of what it means to undermine a special relationship by reconsidering four examples I have previously introduced: the Woods case, the Judas case, the Cambridge Spies case and Keller's example of the disloyal son.

Behaviour that undermines a special relationship is perhaps most easily identified in the Woods case. As I explained in section 3, the marriage relationship between Woods and

Nordegren was shaped by constitutive expectations of caring directly for each other, mutual self-disclosure and exclusivity. In being unfaithful and disappointing those normative expectations Woods' behaviour constituted a disloyal affront against Nordegren. But that is not all; it also compromised his special, marital, relationship with her. *Their partnership could not continue as long as he was unfaithful.* My claim is not just that it would be emotionally and personally challenging for Woods and Nordegren to continue their relationship – perhaps because of the feeling of "distance" it might put between them or because of Nordegren's understandable resentment or hostility toward Woods' actions. Rather, my claim is that their relationship could not be sustained as constituting *an exclusive relationship between partners* as long as he was cheating. Even if Woods and Nordegren did not get divorced and so were technically still married, as long as Woods kept cheating, their relationship would not be one characterised by the constitutive expectations of marriage-like relations. Similar interpretations can be given of the Judas case and Cambridge Spies case. In selling Jesus to those who wanted to kill him, Judas was not merely doing an affront to Jesus. Rather, he compromised the very friendship between them. Without some type of restitution and forgiveness the special relationship between Judas and Jesus would not be able to continue. And in their treachery the Cambridge Spies do an affront against Britain and her people that compromises the continuation of their special relationship with them. Unless their punishment involves being banished from the land, the spies may remain in a citizen-nation relation of sorts with Britain. They might not lose their citizenship, but presumably they would lose some of the benefits and freedoms that go with it. The government will not continue to treat them the same as it does other citizens. It may still be expected to uphold its human rights obligations toward the spies but it will not do so in virtue of their being citizens whom it is obligated to care for. In their treachery they have forfeited the benefits of that special relation.

In contrast to the above cases, consider again Keller's example of the disloyal son who fails to give his jacket to his mother. The son's behaviour constitutes an affront against the mother. Given their special relationship he is obliged to care for her and in keeping his jacket for himself he fails to do so. But that affront does not undermine the son's relationship with his mother. The son may still be disposed to care for his mother and in fact exercise great care for her in other interactions between them. This one failure of his is an affront but does not call the whole relationship into question in the way that the actions of Woods, Judas and the Cambridge Spies do.

Understanding betrayal as a failure to uphold normative expectations in a way that undermines special relationships explains how betrayal is a type of disloyalty while also distinguishing it from mere disloyalty. It also accounts for the features of harm, using a relationship for one's own gain, and deception that can occur in instances of betrayal. On my view, those features can just be seen as part of carrying out an act of disloyalty which undermines special relationships. Undermining a relationship may require deception in order to be successful; it probably will result in harm to the other; and it may be done to benefit one's own interests. Further it can be done deliberately, but it need not be done in that way. The Cambridge Spies undermine their relation with Britain deliberately while Woods undermines his relation with Nordegren negligently.

In section 3 I discussed Jackson's view that Willoughby's violation of his relationship with Marianne in *Sense and Sensibility* was an instance of abandonment rather than betrayal. According to Jackson, betrayal involves viewing a relationship with another as something to be manipulated and Willoughby does not seem to have done that. I agree with Jackson that Willoughby does not see his relationship with Marianne instrumentally. He really does care for Marianne but his behaviour nonetheless undermines a special "pre-engagement" relationship between them. This is evident when he tells his side of the story to Marianne's sister Elinor. Elinor asks whether Willoughby ever took himself to be in a special relationship with Marianne and he replies,

> To have resisted such attractions, to have withstood such tenderness! – Is there a man on earth who could have done it! – Yes, I found myself, by insensible degrees, sincerely fond of her; and the happiest hours of my life were what I spent with her, when I felt my intentions were strictly honourable, and my feelings blameless. Even *then*, however, when fully determined on paying my addresses to her, I allowed myself most improperly to put off, from day to day, the moment of doing it, from an unwillingness to enter into an engagement while my circumstances were so greatly embarrassed. (Austen, 1992, pp. 216-217. Emphasis is Austen's)

Willoughby does not mean to manipulate Marianne but he does undermine their relationship. On my view he does not just abandon her but rather lets himself betray her through poorly made choices.

My understanding of betrayal as a type of disloyalty does not make any assumptions about the moral status of betrayal. In the next section I focus on the morality of betrayal and develop my second argument: while there is always good reason not to betray, not all betrayers are blameworthy.

Section 5: The Morality of Betrayal

All persons who betray are morally responsible for their actions, but not all betrayal is blameworthy. Rather, in some circumstances betrayers can be excused for their acts of betrayal. As I explained in section 3, constitutive expectations within relational domains assume a context of moral responsibility – that those involved in the relationship can be held accountable for failure to uphold such expectations. If someone is unable to be held to account, relevant constitutive expectations will likely not apply to them. For example, we may talk about dogs as a person’s “best friend” and they may behave in ways consistent with some of the constitutive expectations of friendship. They at least appear to desire taking part in joint pursuits with humans. If nothing else they seem pleased to walk along side us and sleep near us. But to hold them to account for failing to uphold the constitutive expectations of friendship would be to attribute to them an understanding of human social norms and motives that I assume dogs do not have. Given that assumption, relevant constitutive expectations of human friendship do not apply to dogs because they are not morally responsible beings. My claim is not just that dogs and other beings that are not morally responsible would not be responsible for acts of disloyalty and, more specifically, betrayal. Rather it is that they could not be disloyal or specifically betray in the first place. Fundamentally, betrayal and other types of disloyalty are a matter of failing to uphold normative expectations one is accountable for meeting. But dogs are not accountable in that way.[50]

My understanding of betrayers as morally responsible does not assume deliberate action on their part. Rather, one can be held to account for things done deliberately and for those done negligently. This can be seen in Woods' betrayal of his wife. As I explained in section 3.2, Woods does not go out meaning to be unfaithful to Nordegren, but he is still accountable for violating his obligation to his wife. I think Keller describes the way in which Woods is morally responsible when, after explaining moral responsibility for deliberate action he writes,

> But you can also be morally responsible for an act when you do not know what you are doing, but you should – when you would know what you were doing were it not for your culpable laziness, self-absorption or distraction; and you can be morally

[50] It is common to talk about dogs as "loyal." My claim that dogs cannot be disloyal or betray need not mean that such talk is misplaced. Rather, as Keller points out, some types of beings such as dogs can be capable of loyalty but not disloyalty. "If you have taken your dog to the park, and he is playing with a large crowd of other dogs and their owners, and when you wander away he notices and runs after you, then he may well be acting out of loyalty; he knows who you are, and he cares about being with you. If, alternatively, he ignores you and trots off after some stranger when the game is finished, then he is not being loyal – but neither is he being disloyal" (Keller, 2007, p. 204). In following the owner, the dog may express loyalty; she is motivated by some special relation between her and her owner. But the dog is not obligated to uphold that special relation. Consider a dog that tends to behave toward its owner in ways that express sympathy to us. When the owner is sad, the dog lays its head on the owners lap or tenderly licks his cheek. If there is a time when the owner is sad and the dog behaves in ways that would express indifference toward him – perhaps the dog simply walks away from the owner and takes a nap – the owner may feel the dog has been disloyal toward him. And as in Keller's example, the dog would be behaving non-loyally. But the dog has not failed to meet an expectation it is obligated to uphold so it is not being disloyal. For it to be so obligated it would need to be capable of understanding and responding, not just to its owner, but to norms governing relations involving humans. I do not think that is something dogs are capable of. So, while I think talk about dogs being loyal is not misguided, talk about dogs being 'disloyal' or 'betraying' should be understood as metaphor.

> responsible for an act when the reason why you do not know what you are doing is that you yourself are denying or suppressing that knowledge. (Keller, 2007, p. 205)

It is appropriate to hold Judas and the Cambridge Spies to account for their deliberate acts of betrayal and to hold Woods to account for his negligent betrayal. All of those betrayers violate normative, constitutive expectations for which they are accountable but Judas and the spies do so deliberately while Wood acts negligently.

That betrayers are morally responsible also leaves open whether their acts are blameworthy or not. I adopt Susan Wolf's understanding of blameworthy action as that which is done when there are "good and sufficient" reasons to do otherwise (Wolf, 1980, p. 159). Given this understanding of blameworthiness, each betrayer in the cases I introduced in section 2 is blameworthy for their actions. The normative expectations shaping the domains of friendship, marriage and citizen-nation relations provide sufficient reason for Judas, Woods and the Cambridge Spies not to betray their respective victims. Also, the value of vulnerable interdependence provides reason not to betray. As Baier has explained, we need to count on each other for the care of things that matter to us: "Given that I cannot myself guard my stamp collection at all times, nor take my rubber tree with me on my travels, the custody of these things that matter to me must often be transferred to others, presumably to others I trust" (Baier, 1986, p. 231). If everyone betrayed those counting on them, our ability to have valuable things cared for would be severely diminished.

Because of constitutive expectations and the value of counting on each other, there are good reasons not to betray. All else being equal then, betrayal is blameworthy. However, there are some situations where betrayal can be excused. I will support this claim by first explaining that not all disloyalty is wrong. I will then extend that explanation to cases of disloyalty that specifically constitute betrayal. Consider the case of an ethical whistle-blower. As Keller points out, whistle-blowing occurs where there is a normative expectation that employees will be loyal to an employer (Keller, 2007, p. 215). Despite that expectation, whistle-blowers can have good reasons for acting disloyally to their employers. Keeping a company secret may mean endangering persons in or outside of the company or it might mean breaking laws. Out of care for persons and respect for the law, whistle-blowers can have reason to not keep company secrets and to violate the norms shaping the employer-employee relationship. But even when the whistle-blower has good reason for going against an employer, her actions still amount to disloyalty. After explaining that ethical whistle-blowers have reasons to act as they do, Keller writes,

> That is not to say that she is not disloyal to the company when she takes its secrets to the public – she is still subject to the expectation, grounded in the employer-employee relationship, that she will keep secrets – but it is to say that there is really nothing wrong with her disloyalty. (Keller, 2007, p. 215)

Keller's claim that disloyalty is not always wrong is applicable to actions constituting betrayal. Despite relevant constitutive expectations and the value of vulnerable interdependence, there can be good reasons to carry out an act of betrayal. When relationships have become corrupt – i.e. when the goal of a relationship or the motivations for sustaining it are bad – and betrayal is employed to resist that corruption, it can be excused. For example, consider the betrayal of the German Democratic Republic (GDR) government by Stasi operative Gerd Wiesler depicted in the film *The Lives of Others* (Donnersmarck, 2006). Wiesler violates his duties as a Stasi operative for the sake of protecting the playwright Dreyman and his partner Christa-Maria Sieland. Dreyman and Sieland are suspected, and are in fact, guilty of treason against the GDR. Wiesler is assigned to observe and report on their actions but he aids the couple by falsifying his reports. Wiesler is disloyal insofar as he fails to uphold normative expectations grounded in his special relation to the GDR as an intelligence officer. Further, his disloyalty constitutes betrayal since his behaviour undermines the relationship between him and Stasi leaders in the GDR. Presumably the main purpose of Wiesler’s role as a Stasi officer is to provide accurate information about citizens to the government so they can monitor and control them. But in falsifying documents Wiesler undermines the special relation he has to the GDR. Whether leaders in the GDR know it or not, he ceases to take part in their machine for surveillance and control.

While Wiesler’s actions constitute betrayal, his blameworthiness is excusable because the normative expectations governing his relation to the Stasi are morally reprehensible. That betrayal can be, all things considered, excusable does not negate the reasons counting against betraying others. Reasons to uphold constitutive expectations and to encourage vulnerable interdependence persist, but acting contrary to those things – undermining a corrupt or immoral relationship – can be excusable in certain circumstances.[51]

To recapitulate, betrayal is best understood as a type of disloyalty that undermines special relationships. It can be done deliberately or negligently and it is blameworthy but can be excused in some situations. Given this understanding of betrayal I now turn to reconsider whether the concept of betrayal can be used to explain the difference between trust and mere reliance.

Section 6: Testing and Explaining Trust’s Vulnerability to Betrayal

I am now in a position to test the claim that trust is vulnerable to betrayal while mere reliance is not. As I will show, both trusters and those merely relying can be betrayed but there is still reason to associate vulnerability to betrayal with trust rather than mere reliance. While trusters can be vulnerable to betrayal in virtue of their trust, features beyond mere reliance must be added to make those merely relying, vulnerable to betrayal.

Accepting vulnerability to someone and trusting them can make one vulnerable to

[51] That betrayal can be excusable when a relationship is “bad” coheres with Baier’s claim that “trust-busting” is not always wrong. “There are immoral as well as moral trust relationships, and trust-busting can be a morally proper goal” (Baier, 1986, p. 232). In section 6 of this chapter I will show that it is not only relationships of trust that can be undermined by betrayal. Rather, persons merely relying on another, or others, can also be vulnerable to betrayal. So excusable betrayal need not only bust trust specifically. It might also bust ‘bad’ relationships of mere reliance.

betrayal. This is because betrayal is disloyalty that undermines special relationships, and trust can establish a special relationship that is liable to being undermined. Even if no previous special relationship exists, trusters can establish one between themselves and another by trusting them. For example, recall the case of the Peruvian artist I presented in *Chapter 1*. The artist entrusted two patrons who were strangers to him; one with the care of his art stall and the other to guide him in his delivery of his art to their home. He was not previously in any kind of special relationship with the patrons. But by accepting vulnerability to the patrons and trusting them, the artist entered into a special relationship with them; one involving normative expectations and obligations. Whether anyone was to predict that the patrons would harm the artist and his stall or not, given the trust relationship between them, the patrons ought not to harm the artist or his stall. If one of them leads the artist astray on the way to deliver the artwork she would be doing him an affront. And if the other patron were to steal items from the art stall while the artist was away he would be doing the artist an affront. If those affronts undermine the special, trusting relationship between the artist and the patrons, they will constitute betrayals. I think that such affronts *do* undermine the trust relationship. If the artist were to find out that he had been led astray or that his stall had been robbed or just poorly cared for, the trust relationship between him and the patrons would most likely not be sustained. He may still rely on them out of necessity but I do not think he would trust them.

The special relationship established by accepting vulnerability and trusting someone explains why trusters are vulnerable to having that relationship undermined; why they are vulnerable to betrayal. Those merely relying can also be vulnerable to betrayal but more than a relationship of mere reliance is needed to establish that vulnerability. To show this I will consider a familiar case of mere reliance and then consider what it would take for there to be an act of betrayal in that case.

The story of Kant's neighbours who set their less reliable clocks by his routine daily walks is a case of mere reliance. The neighbours do not count on Kant because they think he has goodwill toward them, because they expect him to be trust-responsive, because they think he has encapsulated their interests into his own, or because he has trustworthy character. Rather, they presumably just count on him because they inferentially predict that he will continue taking a walk at the same time every day.

What would it take for Kant to betray the neighbours who rely on his habits? If he were to take a nap one day instead of taking his walk, they might be disappointed or simply not know what time it is. But I do not think he would be betraying them in that case. Consider an alternate story where Kant might be able to betray his neighbours. If Kant had agreed with his neighbours to walk at a certain time each day but then walked at different times – whether deliberately or not – I think his actions would be closer to those of betrayal. He would be undermining a special relationship with his neighbours that is based on an agreement he has with them. He would be betraying those who are relying on him.

It might be thought that part of the reason Kant's actions in the above alternate story amount to betrayal is because the agreement between Kant and his neighbours makes their relationship one of *trust*. But I do not think that is correct. Kant's neighbours can make their agreement and merely rely on him without trusting. They may have little recourse but to make their agreement with Kant. Perhaps they are suspicious of him but know that their clocks are unreliable and that he is the only person in their community that has the discipline to act with the precision and repetition they require. They make their agreement and enter into a special relationship with Kant; but they do not trust him.

My alternate Kant story is an instance of persons merely relying on another and being vulnerable to betrayal. But that vulnerability does not result from the fact that the neighbours are relying on Kant. Rather, it results from the special relationship established between Kant and them. While it is possible for those merely relying on another to be vulnerable to betrayal, mere reliance itself cannot establish that vulnerability in the way trust can.

Because it is possible for persons merely relying on others to be betrayed the concepts of trust and betrayal are not exclusively associated. Therefore, Baier, Walker, McGeer, Hieronymi, Wright, McLeod and Holton are not exactly correct in claiming that trusters, but not those merely relying, are vulnerable to betrayal. And yet, their claims are understandable. There is reason to associate betrayal with trust rather than mere reliance because trust, but not mere reliance, can establish the special relationships that betrayals undermine.

In this chapter I have analysed the concept of betrayal as a type of disloyalty and explained the non-exclusive connection between betrayal and trust. But I have not yet said anything about the damage that betrayal can inflict on trust and how we should understand trust after betrayal. In *Chapter 5* I explicate the various ways that betrayal can inhibit trusters and provide an account of trust after betrayal.

Chapter 5: Recovering Reasonable Trust After Betrayal

Section 1: Introduction

The risk of betrayal may fade into the background of many of our interactions, but it will feature in the cognitive and affective experiences of victims of betrayal. While many people take part in social interaction with seeming optimism and confidence about trusting others, persons affected by betrayal may be inhibited in their relationships, perhaps due to a reasonable reluctance to take risks, perhaps due to damaged confidence.

Despite having been betrayed, some persons do trust again. Not all victims of betrayal are unable to engage trustingly with others. They may trust friends, counsellors, and even the very person or institution that once betrayed them. Recall the Cole case I discussed in *Chapter 4*. Cheryl Cole took her husband, Ashley, back after he cheated on her in 2008, only to have him cheat on her again in 2010 (Topping, 2010). After being betrayed in 2008, Cheryl might have divorced her husband. But instead she took him back, accepted vulnerability to him and trusted him again.

Some trust after betrayal is likely to be hasty or misplaced. In some cases trusting an unfaithful partner is unwise. But, presumably, not all trust after betrayal is unreasonable. An explanation is required, then, of how it is possible to recover reasonable trust after betrayal. In this chapter I develop an account of the damage that betrayal can inflict and then provide an explanation of how, given that damage, reasonable trust after betrayal is possible.

In section 2 I present three cases of betrayed trust. Drawing on these cases I identify three main effects that betrayal can have on trusters: distrust; loss of confidence as a judge of trustability; and negative emotional responses. These effects motivate my account in section 3 of the damages that betrayal can inflict. In this chapter I am specifically concerned with trusters, even though those who rely can also be betrayed, because betrayal is more closely associated with trust.

In section 3 I show that betrayal can damage the reasons its victims have for trusting both those who betrayed them and others more broadly. In section 3.1 I analyse the kind of distrust that can be experienced by victims of betrayal in order to explain the damage betrayal can do to the reasons victims have for trusting their betrayers. I show that distrust involves both cognitive and affective elements. It involves a (well grounded) judgement that the betrayer is untrustable – is not competent or committed to doing what they are counted on to do. Distrust also involves an affective attitude of pessimism about another’s trustability. This attitude makes distrust self-confirming, because it shapes the information available to the one who has been betrayed and hence can inhibit them from perceiving good reasons (where such reasons obtain) for coming to trust their betrayers again.

The cases discussed in section 2 show that betrayal can result in distrust not only of one’s betrayers but of others as well. Victims can have what I refer to as “local, collateral distrust” of a specific group of others, or “global, collateral distrust” where they resist trusting all others. In section 3.2, I develop an account of this collateral damage. I follow Margaret Walker and Karen Jones in explaining betrayal’s collateral damage as damage to its victims’ “default trust” (Walker, 2006, p. 90), “basal security” (K. Jones, 2004, pp. 7-16), and “metatrust” (K. Jones, 2004, pp. 7-16). Betrayal damages its victims’ default expectation that others will act as they ought to; it makes them more aware of their own vulnerability; and it causes them to lose trust in trust. I explain that these damages constitute damage to the moral agency of victims of betrayal. Instead of confidently interacting with others, they may be self-protective and restricted in their ability to renew, deepen or establish relationships.

Given these kinds of damages, it looks very difficult, if not impossible, to recover from betrayal. Yet, it is possible to recover reasonable trust. I explore this possibility in section 4. In section 4.1 I explain how it is possible to recover reasons to trust one’s betrayer. The attitude of pessimism which makes distrust self-confirming must be mitigated before victims can come to trust their offenders, should it be reasonable for them to do so. The process of forgiveness is one way that pessimism about trusting another can be mitigated, and optimism awakened. To show this, I argue that forgiveness is best understood in terms of Strawsonian co-reactive exchange and I explicate how the offender remorse involved in such exchange can mitigate victim pessimism. When an offending betrayer reacts to victim resentment by taking responsibility and expressing remorse, the victim can come to forgive the offender and begin to have some optimism about trusting her. As I explain, coming to forgive the offender and being optimistic about trusting her does not guarantee that the victim will come to trust the offender again; but it helps the victim to see features o the offender that might make it reasonable for him to do so. However, the offender remorse that is integral to the process of forgiveness is not always available. The offender may be unavailable to engage in co-reactive exchange with the victim, it may be inappropriate for the victim to interact with the offender, or the offender may be beyond remorse. When offender remorse is unavailable, the recovery of reasonable trust in one’s betrayer through a co-reactive process of forgiveness is inhibited. In such cases, victims of betrayal can, nevertheless, still recover reasonable trust in people other than their betrayers.

In section 4.2 I explain ways in which reasonable trust can recover from more generalised damage. I begin by explaining how victims can recover reasonable trust in those towards whom they feel local, collateral distrust. I argue that victims can do so when they receive enough evidence that not all those in the group they distrust are incompetent and uncommitted to doing what they might have trusted them to do. I then explain how persons experiencing global, collateral distrust can trust others reasonably again. I consider three ways in which their trust can be scaffolded by members of the moral community around them: individuals and institutions can affirm that what was done to the victim was in fact wrong; they can show that not all individuals and institutions are betrayers insofar as they uphold normative expectations governing their relationships; and they can assist victims to regain confidence in their judgements of trustability by modelling good judgments themselves. None of my arguments in section 4.2 require victims to first recover trust in their respective betrayers. Even if resolution in that relationship has not been reached, victims can still recover reasonable trust in people other than their betrayers. I conclude section 4.2, and this chapter as a whole, by identifying an important characteristic of trust identified by my analysis. The recovery of reasonable trust after betrayal must be done with others. Whether we are engaging in co-reactive exchange with our betrayers, receiving evidence about those toward whom we have local, collateral distrust, or having our trust scaffolded by the moral community, we cannot deal with the damage that betrayal inflicts on our own. Trust is relational even when it is being repaired.

Section 2: Three Cases of Betrayed Trust

The cases in this section show that betrayal can affect its victims in at least three ways: they may distrust their betrayers as well as those who have not betrayed them; lose confidence in their own judgements of others' trustability; and have negative emotional responses such as strong anger and hostility.

Major Damage to the Trust of Minors

In April 2010 Cameron Tweeddale Smith pleaded guilty to 11 charges of committing an indecent act in the presence of a child under 16 and committing an indecent act in the presence of a child under 16 while under care or supervision. These offenses took place while Smith worked as a first-aid attendant at Brighton Grammar School in Victoria, Australia. Given that he was an adult and in a role as a school staff member and health care worker, it is plausible that the children Smith victimized inhabited a climate of trust and saw him as trustable. This perception of Smith is evident in the testimony of one abuse survivor who said about Smith, “I trusted him and thought he cared about me” (Lowe, 2010). I think this survivor’s testimony should be taken at face value. Smith’s actions were, prime facie, a betrayal of trust.

Smith’s betrayal of his victims’ trust resulted in the victims acquiring an attitude of distrust toward him and toward others. As one boy expressed in a victim impact statement, “The eventual betrayal of trust has made me reluctant to trust professionals” (Lowe, 2010). I will refer to the type of distrust that persons can have toward those who did not directly offend them as “collateral” distrust. Such distrust is distinct from “direct” distrust of an offender. Direct and collateral distrust are not unique to the Smith case. The next case of betrayed trust also results in direct and collateral distrust but, in addition, it involves a loss of confidence in oneself as a judge of trustability and a deep sense of anger.

Betrayed by a Trusted Priest and Church

Recall the case of sexual abuse that I discussed in *Chapter 3* and *Chapter 4*. In that case, the abuse survivor was involved in the Catholic Church from an early age, and then was severely violated by a priest whom she had known for many years. Like the Smith molestation case above, the Catholic Church case is, prime facie, an instance of betrayed trust. Further, as I explained in *Chapter 4*, the survivor’s trust was not only violated by the priest who directly abused her. Both the priest and the church that oversaw his role – and should have held him to account for his actions – failed to uphold the normative expectations governing the special relationship between priest and parishioner.

This abuse survivor recounts how the priest’s betrayal resulted in her experiencing not only direct but also collateral distrust. “I have lost all confidence in the Church’s leadership. I distrust most clergy and suspect most of what they preach is not authentic. That’s not to say that it is, it's just how I feel when I hear them preach" (Fleming, 2007, p. 165). While the resulting distrust in this case is similar to that experienced by the abuse survivor in the Smith molestation case above, I think these words point to an additional effect of betrayal on trusters. The survivor does not simply say that she distrusts most clergy; she says that she suspects that what they preach is inauthentic and admits that her suspicion may be inaccurate. A questioning lack of confidence about who can be trusted has worked its way into her psyche. What this testimony shows, therefore, is how betrayal can give rise to a loss of confidence on the part of trusters about their capacity to reliably judge trustability. If the survivor had previously taken herself to have good reason to trust priests and the Catholic Church only to have her trust betrayed, I think it would be understandable for her to begin to doubt the accuracy of her judgements about trustability.

In addition to her distrust and loss of confidence, the victim of priest abuse experiences deep anger at being betrayed. This is evident when she says,

> Facing my anger was difficult. My family dynamic did not include expressing anger. I had learned from an early age to repress my anger. I learned to pout instead of being angry. My expression of anger over my abuse had been muted. My therapist encouraged me to write what I was angry about. She encouraged me to hit a pillow with a Wiffle ball bat, or punch a pillow or scream. All in safe places. At times the hurt of it all overcomes me and I have to take out a pen and a paper and write why I am angry. The list usually looks like this: I'm angry because I was used. I'm angry because I was abused. I'm angry because I felt powerless. I'm angry because I allowed it to happen. I'm angry because he took away precious years from me, years when I could have fallen in love and begun a family of my own. I'm angry because I placed so much trust in him and that trust was betrayed. I'm angry at the manipulation and control. I'm angry because I did so much for him and he repaid it with his abusiveness. I'm angry because it has disrupted my life so much. I'm angry at the Church for not taking responsibility for its clergy. I'm angry at the loss I've experienced in my Catholic faith. I'm angry because of the pain. I'm angry because I enabled him to be successful in his endeavours. I'm angry because he blames me. I'm angry at his denial. I'm angry that I lived such a lie for so long. I'm angry at the institutional Church for being so incredibly dysfunctional. (Fleming, 2007, p. 152)

This case shows that betrayal can result in direct and collateral distrust; it can affect a truster's confidence about judging trustability; and it can leave the betrayed with strong negative emotional responses. The next case shows that these effects of betrayal do not feature only in the experiences of persons who have been betrayed themselves. Betrayal can affect third parties. That is, in addition to affecting its “primary victims”, betrayal can also affect those who are its “secondary victims.”[52]

No One Left to Trust

For two years Russell Williams, a former colonel in the Canadian Forces, carried out murders, sexual assaults and break-ins. One of his victims was Jessica Lloyd, a 27 year old woman who vanished in January of 2010. After Lloyd’s disappearance, Williams was caught and brought to trial. He was found guilty on two counts of first-degree murder, two counts of sexual assault, two counts of forcible confinement and 82 charges of break-and-enter and attempted break-and-enter (Nguyen, 2010).

During Williams’ court hearing, family and friends of Lloyd spoke about their late friend’s life and her killer’s horrendous offense. One of Lloyd’s friends, identified only as “Lisa”, gave the following testimony:

> The past nine months has been an emotional roller coaster for me, it has gone from hope and faith to shock and betrayal and now anger, hate and depression. A new side of me has been brought out that I had never seen before . . . I constantly ask myself “Why? Why Jess?” I have a feeling of guilt wondering why she was chosen and why I’m still here. My faith in God has completely diminished. If there is truly a God out there, how could you let this happen to such an incredible person? Or better yet, how could you create a human to do such a monstrous thing to another?
>
> My sense of trust is now gone. I now trust no one, whether it is a stranger, neighbour or even someone in uniform.
>
> Throughout . . . I was taught at a young age to be proud and respectful around authorities. But the very person whose job it was to protect my country and keep me safe was actually terrorizing my local community making all females live with the fear of being attacked and Jess’s life was taken along his path of destruction. This leaves me with a question, “Who can I trust?” No one. (Nguyen, 2010)

I think Lisa’s talk about betrayal and undermined trust should be taken at face value. As I will explain, while I do not think the interactions between Williams and his victims

[52] I follow Trudy Govier and Margaret Walker in my use of “primary victims” and “secondary victims” (Govier, 2002, pp. 92-95; Walker, 2006, p. 163). Primary victims are those who have been the direct target of another’s actions while secondary victims are those who have been affected by actions that are targeted against others.

should necessarily be understood as instances of trust, Williams' actions do constitute betrayal. And as is evident in Lisa's testimony, the betrayal perpetrated by Williams affected the trust of those close to at least one of his victims.

The interactions between Williams and his victims are not characterised by trust. As I explained in *Chapter 1,* to count as an instance of trust an interaction or relationship must be sufficiently similar to paradigm cases of trust, which involve counting on someone and accepting vulnerability to practical and moral harm. Williams' victims were vulnerable to him but it would be inaccurate to say that they counted on him or accepted vulnerability to him either explicitly or implicitly. They did not assess the risks of interacting with Williams and then explicitly accept the vulnerability involved in doing so. Nor did they implicitly accept vulnerability to him by gradually developing a relationship with him over time. Until he broke into their lives and abused them, they had no interaction with him at all.

However, while the interactions between Williams and his victims were not characterised by trust, his actions against them nevertheless constitute betrayal. In *Chapter 4* I argued that betrayal is best explained as a type of disloyalty that undermines special relationships by violating normative expectations in a way that compromises the ability of those relationships to be sustained. Williams' actions violated a normative expectation that he would, qua soldier, protect his country and its citizens. As Lisa testified, "...the very person whose job it was to protect my country and keep me safe was actually terrorizing my local community making all females live with the fear of being attacked..." (Nguyen, 2010). Further, Williams' actions violated relevant normative expectations in ways that undermined the citizen-soldier relation between Williams and his victims. As long as he persisted in his offenses, Williams would not be fulfilling his role as a protector of Canadian citizens. Insofar as Williams' offenses violate a normative expectation in a way that undermines a special relation, they constitute betrayal.

Lloyd's friend Lisa was not his primary victim and yet distrust and negative emotional responses are evident in her testimony. She speaks of losing her "sense of trust" (Nguyen, 2010); of trusting no one and feeling hate. Presumably, Lisa would include Williams in the universal group of persons she can no longer trust. And so her distrust is both local and collateral. She has direct distrust of the offending betrayer but she also has extensive collateral distrust of others. It is this collateral distrust that is affecting her life.

Lisa also appears to experience some loss of confidence in her judgements of trustability. Although this isn't a case of trusting an individual, only to be proven wrong, Williams' actions upset assumptions about persons in authority. As she said,

> I was taught at a young age to be proud and respectful around authorities. But the very person whose job it was to protect my country and keep me safe was actually terrorizing my local community making all females live with the fear of being attacked and Jess's life was taken along his path of destruction. (Nguyen, 2010)

In other words, Lisa questions her judgements about who should be trusted as well as the assumptions influencing those judgements. Lisa's distrust, negative emotional responses, and possible loss of confidence as a judge of trustability are important because they show that betrayal does not only have primary victims; there are also secondary victims of betrayal.

The above cases show that betrayal can generate direct and collateral distrust, undermine a person's confidence in their judgements of trustability, and trigger negative emotional responses in persons whether they were the primary or secondary victims of betrayal. In the next section I analyse these effects in order to identify precisely what needs to be repaired for victims of betrayal to place trust reasonably again.

Section 3: An Account of the Damages That Betrayal Can Inflict

In addition to undermining or destroying the reasons its (primary and secondary)[53] victims had for trusting those who betrayed them, betrayal damages the reasons its victims had for trusting others. In this section I consider why betrayal would result in those effects. I begin with betrayal's direct damage.

Section 3.1: Betrayal's Direct Damages

Betrayal damages the reasons its victims had for placing trust in those individuals, role-holders or institutions that betrayed them. This is evident in the Smith molestation case and in the Catholic Church case, where each victim distrusts their betrayer(s), for good reason. The survivor in the Smith molestation case says of Smith, "I trusted him and thought he cared about me" (Lowe, 2010). And after her betrayal, the survivor in the Catholic Church case distrusts the offending priest who abused her and the institution that protected him. I will now explicate this damage by analysing the concept of distrust.

Accounts of distrust have been developed by Hardin (Hardin, 2002, 2004), Govier (Govier, 1992a, 1992b, 1994) and Jones (K. Jones, 1996).[54] From Hardin I develop an understanding of distrust as involving a cognitive assessment of another as untrustable. Then, using Govier's view, I show that a cognitive assessment on its own cannot account for distrust's self-confirming nature. Finally, I use Jones' view of distrust as an affective attitude of pessimism about another's goodwill to identify an additional affective component involved in distrust. In the end, I understand distrust to be a resistance to trust with cognitive and

[53] For the remainder of this chapter I will simply refer to "victims" of betrayal rather than each time qualifying that the victim could be primary or secondary. I will only distinguish between the two when doing so is important for explicating a kind of damage that betrayal can do.

[54] Others, such as Baier (Baier, 1986, 1994), Potter (Potter, 2002), O'Neill (O'Neill, 2002a, 2002b), and McGeer (McGeer, 2002) have mentioned distrust when analysing the concept of trust, but they have provided little explication of the concept of distrust itself. Baier and Potter, at least, imply understandings of distrust in their accounts of trust. I will briefly explain these implied views in footnotes when developing my analysis of the cognitive assessment involved in distrust.

affective components.[55]

Hardin takes distrust to be a cognitive assessment that another is untrustworthy (Hardin, 2002, p. 89; 2004, pp. 3, 9): “If I distrust you, that is because I think that your interests oppose my own and that you will not take my interests into account in your actions” (Hardin, 2004, p. 11). According to Hardin, the trustworthy are competent and committed to doing what we are counting on them to do because doing so is in their interests (Hardin, 1996, p. 27; 2002, p. 35). But as I showed in *Chapter 2*, this is better understood as trustability than trustworthiness. Likewise, Hardin’s view of distrust amounts to a cognitive assessment that another is not trustable on the basis of their interests: distrust is a matter of thinking that another will not encapsulate my interests into their own[56] (Hardin, 2004, p. 11).

Hardin is right that distrust can involve cognitive assessment. Recall the example of mere reliance I presented in *Chapter 1* involving Victorian workers. While some workers purchased adulterated goods without knowing it, some would have doubted the integrity of the food the vendors were selling but purchased it regardless because they had no choice. Such workers might have relied on the dishonest vendors without trusting them. And, presumably, some of them did not only *fail to trust* the vendors but rather *distrusted* the vendors while still relying on them for food. One of the reasons the vendors were not to be trusted was perhaps that it was not in their interests to provide only good quality merchandise. There was, presumably, little the workers could have done to hold the vendors to account for their products. The vendors had little to lose, and perhaps much to gain in

[55] In *Chapter 1* I followed Govier and Jones in holding that trust and distrust are contrary but not contradictory (Govier, 1992b, p. 18; K. Jones, 1996, p. 15). They are mutually exclusive but not collectively exhaustive (Ullmann-Margalit, 2004, p. 60). In addition to trusting or distrusting another, I may simply fail to trust them. For example, I might merely rely on a bus driver to get me to work rather than trusting her to do so. My interaction with the bus driver is not characterised by trust; and yet I would not say that I distrust her. I accept that distrust is distinct from a failure to trust, or lack of trust. However, I leave that distinction to one side as it is peripheral to my argument that betrayal damages reasons its victims have for trusting others.

[56] Deborah Welch Larson and Ullmann-Margalit assume similar, interest-based views of distrust. Larson writes, “Distrust is a judgment that one cannot depend on the other’s actions or promise because the other has an interest in cheating or wishes to harm oneself” (Larson, 2004, p. 35). And Ullmann-Margalit says, “I begin to distrust you when I am in a position to form the actual belief that you do not intend to act in my best interests in that matter” (Ullmann-Margalit, 2004, p. 61).

adulterating their goods.

While distrust can involve a judgement that another will disregard my interests, there can be other reasons for distrust. I will show this by explaining what it means to be *un*trustable rather than trustable. As I explained in *Chapter 2*, the term "trustability" refers to the attribute of being one whom another has reason to trust. It can be reasonable to trust others because they are competent and committed to doing what you are counting on them to do. And they can be committed to doing so for a number of reasons. I may think another is committed to doing what I am counting on him to do because she is motivated to encapsulate my interests into her own, or because she has goodwill toward me, is (or will be) trust-responsive to me, or because she has virtuous, trustworthy character. Similarly, I may have reason to withhold my trust from another because I take her to be relevantly incompetent or uncommitted to doing what I would have counted on her to do. And I might think she is uncommitted because I do not think she is motivated to encapsulate my interests; because I think she has ill will toward me[57]; because I do not think she would be responsive to my trust; or because I do not think she is a trustworthy person – she lacks integrated, virtuous character.[58]

At the start of this section I assumed that the victims in the cases of betrayal discussed in section 2 had good reason to distrust their betrayers. My explication of distrust as involving a judgement that another is untrustable explains some of these reasons. Through their acts of betrayal, the betrayers have shown that they have not encapsulated their victims'

[57] Although Baier does not explicitly analyse the concept of distrust, an understanding of distrust as involving an assessment that another has ill will toward oneself is implied in her will-based view of trust. As I explained in chapter 1, Baier understands trust to be reliance on another's goodwill (Baier, 1986, p. 234). Given that view of trust, it follows that if you refrained from relying on another because you thought she had ill will towards you, you would be distrusting her.

[58] An understanding of distrust as involving an assessment that another does not have virtuous, trustworthy character is implied in Potter's virtue-based view of trust (Potter, 2002, p. 16). Potter understands trust in the following way: "A trusts B to be x sort of person with regard to y, where 'x' = (from A's perspective) a positive quality of character or way of performing an action and where 'y' = some good that A values" (Potter, 2002, p. 17). Given this view, distrust would involve refraining from trusting someone with regard to something *specifically because one takes the other to lack virtuous, trustworthy character.*

interests into their own; they have expressed ill will toward their victims; they have responded negatively to their trust; and they have shown themselves not to have integrated, virtuous character. Victims of betrayal thus have good reason to judge their offenders to be untrustable.

However, while distrust can involve a cognitive judgement, it can also involve an affective component. Specifically, distrust can involve pessimism about another's characteristics and behaviour.

Govier points out that distrust can shape the way we perceive others: "When we distrust a person, even evidence of positive behavior and intentions on his or her part is likely to be received with suspicion, to be interpreted as misleading and, when properly understood, as negative after all" (Govier, 1992a, p. 52). Elsewhere Govier shows that distrust can be self-confirming: when we distrust others, "...we interpret their further actions and statements consistently with these negative expectations (Govier, 1992b, pp. 17-18).

Govier does not explain why distrust influences our perception of those we distrust. However, Jones' view of distrust can explicate the interpretative dimensions of distrust. In keeping with her view of trust as a quasi-perceptual affective attitude, Jones understands distrust as involving an affective attitude of pessimism about another's goodwill and competence (K. Jones, 1996, p. 7). As I explained in *Chapter 1*, in Jones' view, the affective attitude of optimism involved in trust makes it like a quasi-perceptual emotion. Trust is not just a response to the world; it also affects how we see the world. It draws our attention away from certain features of the world and toward others. Jones also understands distrust to be quasi-perceptual. It is because distrust is quasi-perceptual that it is self-confirming. Because the affective attitude involved in distrust disposes persons to see the world as fitting to their pessimism, distrust can make persons see others as deserving distrust – as untrustable.

According to Jones, the attitude of pessimism is directed at the other's will. I think it can be understood more broadly. In distrusting another, an individual could be pessimistic about the other's interests; their will towards them; their trust-responsiveness; or their character. For example, I think it would be understandable if the survivor in the Catholic Church case was pessimistic about the priest having good will toward her; encapsulating her interests into his own; being favourably responsive to her should she count on him; or having integrated, virtuous character. His violation may even give her good reason to be pessimistic about *all* of these things. The attitude of pessimism involved in distrust is thus more general than Jones suggests: distrust involves an attitude of pessimism about another's characteristics and behaviour.

There is a bidirectional relation of influence between the cognitive and affective components involved in distrust. My judgement that you are relevantly incompetent or uncommitted to doing what I am counting on you to do can lead me to be pessimistic about your characteristics and behaviour. And if I have that attitude of pessimism toward you, I am more likely to judge you to be untrustable. Hence, again, distrust is self-confirming.

The judgement and attitude involved in distrust are not coextensive. It may be unlikely that one would judge another to be untrustable and *not* be pessimistic about him. But it is possible to acquire the attitude of pessimism without necessarily making a cognitive judgement about another's trustability. This is the case with distrust which results from betrayal. After being betrayed, the victims in the cases discussed in section 2 need not have stopped to consider their offender's trustability in order to be pessimistic about competence, interests, will, responsiveness or character. Presumably, they would be pessimistic about such things just because they have been betrayed by their respective offenders. Each direct victim has been badly treated by an offender and the secondary victim – Lisa – is aware of her good friend's bad treatment.

Distrust, then is a resistance to trust which can involve a cognitive assessment that another is untrustable and/or an affective attitude of pessimism about another. With this view of distrust in hand, I will now identify two ways that betrayal can damage the reasons its victims had for trusting those who betrayed them. First, it can provide its victims with reason to re-assess their respective betrayers. Insofar as a betrayer fails to uphold a normative expectation governing a special relationship between himself and another, he shows that he is not committed to sustaining that relationship and doing what the other is counting on him to do. For example, by violating his students, Smith shows that he is not committed to sustaining the special relationship between school staff member and student. So part of the way that betrayal damages the reasons its victims had for trusting those who betrayed them is that it shows that those reasons do not obtain: the offending betrayer is not committed to upholding the special relationship between himself and his victim(s).

Second, my analysis of distrust shows that betrayal can constrain the reasons victims of betrayal may have for rebuilding future trust. Because they have reason to judge that their betrayers are not committed to upholding the special relationship that might previously have existed between them, victims may be pessimistic about placing trust in their offenders. And that pessimism may be compounded by negative emotions that the victims might feel towards the betrayer. This pessimism can be quasi-perceptual, thus inhibiting them from perceiving any new information about their betrayers as giving them reason to trust the betrayers again. Distrust can be self-sustaining.

The damage caused by betrayal to the reasons victims of betrayal might have had to trust their betrayers need not always be regarded negatively. It can keep victims of abuse, such as the survivors in the Smith molestation case and the Catholic Church case, safe from placing trust unwisely in their abusers. But the damage inflicted by betrayal can also inhibit its victims from rebuilding a relationship that in some cases they might want to restore. Recall the Woods betrayal case that I discussed in *Chapter 4*. Given that Woods and

Nordegren have children together, Nordegren might have wanted to rebuild her relationship with Woods, if only so for the sake of her children.

To rebuild a relationship broken by betrayal, victims need to find a way to place trust reasonably in their betrayers – which is possible, if the betrayer really does become competent and committed to doing what they had been counted on to do. The offending betrayer needs to prove themselves to be relevantly competent and committed, and the victim's pessimism needs to be mitigated so as to allow them to recognise the new reasons available for trusting their former betrayers. Such a process requires something of both betrayers and their victims. It requires what McGeer calls a series of "co-reactive" exchanges between them (McGeer, 2011). I will discuss the role of co-reactive exchange in recovering reasonable trust in one's betrayer in section 4.1.

Betrayal has effects beyond the immediate relationship between two parties. It can create collateral distrust toward others. The student in the Smith molestation case finds it difficult to trust other professionals in general, not just the one who betrayed him. The survivor in the Catholic Church case experiences collateral distrust towards other clergy. And Lloyd's friend Lisa has global distrust. I now turn to explicate this additional, collateral, damage.

Section 3.2: Betrayal's Collateral Damage

In addition to showing that betrayal has effects beyond the immediate relationships it damages, the cases in section 2 show that betrayal can result in collateral distrust that varies in its extension. The Smith molestation case and the Catholic Church case show that victims of betrayal can distrust a specific group of others – i.e. professionals, or clergy. I will refer to such distrust as "local, collateral distrust." But as Lisa's testimony reveals, collateral distrust resulting from betrayal can extend more broadly. Victims of betrayal may distrust everyone. I will refer to such distrust as "global, collateral distrust." In this section I identify the damage done by betrayal that explains this range of collateral distrust.

Both Walker and Jones discuss the way that damage to trust can generalise. Walker claims that betrayal can damage one's beliefs about oneself, one's judgement about what is "right" and "wrong" and it can damage what she terms "default trust" (Walker, 2006, p. 90). Jones claims that betrayal can damage "metatrust" – that is, our trust of trust – causing us to revise our practices of trust and, when severe, it can damage what she calls "basal security" (K. Jones, 2004, p. 11). Through considering and critiquing these claims made by Walker and Jones I will develop an account of betrayal's collateral damage. I explain global, collateral damage as a result of damage to one's default trust and basal security. Betrayal can lead us to think that others will not uphold relevant normative expectations, and it can cause an awareness of our own vulnerability to feature significantly in our affective experiences. As a result, we may lack self-confidence as a truster and have global, collateral distrust. But that general lack of confidence does not explain why the collateral distrust that some victims have is *local*, rather than *global*. I identify two ways in which local, collateral distrust can come about: it can result from revisions that victims make to their trust practices; and it can arise as an attitude of pessimism that is directed towards those who are associated with one's betrayer.

Walker claims that betrayal can damage one's beliefs about oneself, one's sense of judgement, and the world (Walker, 2006, p. 90). She connects this to betrayal's damage of default trust:

> In the worst case of damaged trust among intimates, however, it is not only the violated relationship that is shattered but a whole nexus of the injured person's beliefs about himself, his judgement, his understanding of a shared history, and even the nature of 'people', 'the world,' and 'right and wrong.' These are cases where individual violations of trust have been due to terrible injuries or stunning betrayals, and therefore may compromise, at least in some ways or for some period of time, not only trustingness in relationships but certain kinds of default trust as well. (Walker, 2006, p. 90)

To explain the damage that Walker has identified, her concept of default trust must be understood.

Walker develops the concept of default trust in order to account for the “trust” that persons can have toward institutions, groups and others in general without identifying particular individuals or role-holders that they are trusting (Walker, 2006, p. 83). Default trust “is that unreflective and often nonspecific expectation that strangers or unknown others may be relied upon to behave in an acceptable and unthreatening manner [...] We trust *that* they will behave as they should” (Walker, 2006, p. 84. Emphasis is Walker's).

Walker argues that this default trust is a kind of trust because persons feel Strawsonian participant reactive attitudes and a sense of being betrayed when it is upset.

> What I am calling default trust might seem like mere comfort or habit, but the occurrence of negative reactions of indignation, resentment, and betrayal when default trust is disrupted or disappointed is the signal that our reliance is trusting, and that it assumes others’ responsibility and not just their predictability. (Walker, 2006, p. 85)

Walker’s argument here is based on her participant stance approach to trust. As I explained in C*hapter 1*, on that approach, trust involves taking a participant stance towards another and holding them responsible to do what they ought to do (Walker, 2006, p. 79). Because those we trust are morally responsible participants, we have reactions of resentment and indignation when they violate our trust. But, as I showed in *Chapter 1*, it is possible to take a participant stance towards another and have reactive attitudes towards her while merely relying. The Victorian worker example, in which a shop vendor shows indifference to the worker who purchases her products, presents such a case. The worker might be prepared to resent the vendor and rely on her merchandise but not think that she is someone to be trusted. Having, or being prepared to have, reactive attitudes towards someone is not sufficient for trust.

While the view of trust on which Walker grounds her understanding of default trust is unsatisfactory as a complete account of trust, default trust can be sufficiently similar to paradigmatic cases of trust to count as a kind of the cluster concept of trust. Recall the paradigmatic "trust fall" example I presented in *Chapter 1*. In that example, the blind-folded participant is vulnerable to her teammates, but accepts vulnerability to them because she thinks that they have goodwill towards her and/or will be responsive to the fact that she needs them to catch her. Default trust can be sufficiently similar to this paradigmatic case to count as a kind of trust. When we have an unreflective background expectation that others will act in ways that they ought to, we are vulnerable to those others taking advantage of that expectation and harming us. A community with a high level of default trust would be a prime location for a confidence trickster.

Despite not involving explicit acceptance of vulnerability to others, as occurs in the "trust fall" case, default trust does involve implicit acceptance of vulnerability. Persons who have default trust towards others unreflectively engage in interaction with them. They may not weigh up the risks of doing so and then voluntarily choose to accept vulnerability to those others. But by interacting with others, persons who have default trust towards others implicitly accept vulnerability to them.

Default trust is also similar to the "trust fall" example insofar as it involves accepting vulnerability because of an expectation about others. The blind-folded participant falls because she takes her teammates to have goodwill towards her or expects them to be trust-responsive. Those with default trust confidently interact with others because they expect those others to behave as they should. Rather than falling into the arms of our teammates, if we have default trust, we ask strangers to watch our bags at the airport, leave the doors to our houses unlocked and allow ourselves to fall asleep on public buses. Given its similarity to the paradigmatic "trust fall" example, I understand default trust to be a kind of trust. It is a background kind of trust characterised by a cognitive expectation that others will act as they ought to – that they will uphold relevant normative expectations. As such, Walker's claim about the damage that betrayal does can be understood to include damage to one's beliefs about oneself, one's judgements, one's understandings about "people" and "the world", and one's normative expectations about others.

Jones also claims that betrayal damages certain background assumptions that are important for trusting others broadly. However, her claim is distinct from Walker's. The basal security that Jones identifies betrayal as damaging is an "affectively laden state" rather than a cognitive expectation (K. Jones, 2004, p. 8). Jones posits the concept of basal security to account for divergence that can occur between a person's assessments of risk and her actual behaviour (K. Jones, 2004, p. 8). To borrow an example from Jones, "A woman might find herself unable to follow through with her plans to go out with friends, though she judges that doing so poses hardly any risk" (K. Jones, 2004, p. 8).

Jones explains basal security as an "affectively laden state" or "set of dispositions" which can be established by nature and nurture and that affects how the world is experienced by individuals (K. Jones, 2004, pp. 8, 9). Because of one's genes, character, personal history and the way one was influenced by parents, carers, or lack of care, individuals can be disposed to interpret themselves as more or less vulnerable to being harmed by others and they can tend to see social interaction as more or less risky (K. Jones, 2004, p. 10).

> Whereas a given degree of risk might not enter at all into the practical deliberation of someone with high basal security, or if it enters it might enter as challenge rather than threat, someone with low basal security lives in continual awareness of her own vulnerability. (K. Jones, 2004, p. 9)

Jones suggests that betrayal can cause its victims to lose "trust of trust" and, when severe, it can damage their basal security (K. Jones, 2004, p. 11). I will follow Jones in understanding the loss of trust in trust that betrayal can cause as a loss of "metatrust" (K. Jones, 2004, p. 11). To regain metatrust, victims may seek some explanation for what happened to them by reconsidering their practices of trust: "We assume partial responsibility for the betrayal and take it to signal some fault or excessive optimism in our trust" (K. Jones, 2004, p. 11). If victims are able to identify excessive optimism or a fault on their part, they may revise their trust practices in an attempt to avoid future betrayals. Such revision can be done with respect to a specific individual or institution, a certain kind of other, or a certain kind of vulnerability: "Thus I might resolve never to tell John my secrets again, never to tell colleagues my secrets, or to be more circumspect in the telling of secrets, period" (K. Jones, 2004, p. 11). The victim redefines the way he trusts so that he distrusts where he used to trust.

When betrayals are severe, and we are unable to explain them by merely re-evaluating our practices of trust, they do not only cause us to lose metatrust. They can also damage our basal security. Jones writes,

> But some betrayals shake our metatrust while leaving no identifiable ways of retrenching on our first-order trust so as to regain trust in our trust. These are the betrayals that, if serious, shatter basal security. If the betrayal is serious, and if I can identify nothing in my first-order trust that was unwise or overly optimistic, then I cannot attribute part of the fault of betrayal to my own practices of trust. I come face to face with my own vulnerability and with the inability of even the wisest of trust practices to protect me from the harm that others can inflict. (K. Jones, 2004, p. 11)

When we have been made aware of how vulnerable we are to betrayal, even when we place trust reasonably, our sense of insecurity may dispose us to resist trusting anyone. Because our basal security has been damaged we have global, collateral distrust.

Neither Walker nor Jones presents an account of why betrayal damages one's default trust and basal security. However, the account of betrayal that I developed in *Chapter 4* is able to explain this. In chapter 4 I argued that betrayal is a kind of disloyalty. I followed Keller in understanding disloyalty as failing to take something's side when you ought to do so given your special relationship to it (Keller, 2007, p. 211). As such, violation of a normative expectation is a central feature of disloyalty. I distinguished betrayal from mere disloyalty by identifying the way in which betrayal's violation of normative expectations undermines relationships.

It is precisely because betrayal involves a violation of normative expectations that it can damage one’s expectation that others will do what they ought to – i.e. one’s default trust. When a person is betrayed, he comes to see that the norms governing the special relationship he had with his betrayer, and the obligations associated with those norms, were not able to protect him from harm. This lack of protection resulting from the violation of normative expectations can also explain why betrayal damages basal security. In being betrayed, the victim of betrayal has been made aware that norms can be broken. And if one person could violate those norms, others might as well. That is why the victim realizes that he is vulnerable: the norms that he had assumed would be upheld have been stripped away. Betrayal can also cause its secondary victims to become aware of their own vulnerability. Lisa might know that her friend did not do anything to enable Williams’ horrendous actions: her friend followed all the norms and principles for living a safe flourishing life that Lisa does. Yet Lisa can now see that, despite those precautions, her friend was not safe. Lisa might realize that she too, in spite of precautions, is vulnerable to harm.

Damage to default trust and damage to basal security, and the sense of vulnerability aroused by this damage, can thus explain Lisa’s global, collateral distrust in the Lloyd murder case. Since Williams failed to uphold normative expectations governing his role as a protector of Canadian citizens and as a member of society himself, he has shown Lisa that the norms she had assumed would be upheld may be violated. Given that realization, Lisa’s own vulnerability to other menacing individuals roaming the community has become all too real to her. Her default trust and her basal security have been damaged. In that state, she has global, collateral distrust – there is no one left that she trusts.

As I have shown, not all collateral distrust is global. After being betrayed by her husband twice, Cheryl Cole may think that “all men are bastards” and distrust them but not have *global* distrust. Presumably, she would still trust women. If her default trust and basal security were damaged by her husband’s betrayal of her, then her distrust would be more extensive. She would self-protectively distrust everyone. But she, presumably, does not. Such localized, collateral distrust requires explanation.

Jones' view of how victims of betrayal can revise their trust practices can explain some kinds of local, collateral distrust. As I explained above, betrayal can damage one's trust in trust (K. Jones, 2004, p. 11). In an attempt to explain what has happened to them and regain metatrust, victims of betrayal may take partial responsibility for the betrayal done to them and revise their trust practices to guard against similar offenses occurring in the future (K. Jones, 2004, p. 11). Deliberate revision of trust practices can explain the Catholic Church survivor's local, collateral distrust of clergy. After saying that she has lost confidence in the Catholic Church's leadership and that she suspects that what clergy say is inauthentic, the survivor says, "I will never follow blindly again. I will discern my own beliefs and live my faith in the way God calls me" (Fleming, 2007, p. 165). She seems to have attributed part of the reason that she was betrayed to previous blind obedience to the priest and church. She has resolved to change her trust practices and discern her own beliefs.

However, not all local collateral distrust is arrived at as a result of deliberate decisions to revise one's trust practices. The victim in the Smith molestation case says, "The eventual betrayal of trust has made me reluctant to trust professionals" (Lowe, 2010). While not explicit, it is implied in his statement that he was passive in coming to distrust professionals. I do not think he deliberately revised his trust practices so that he no longer trusts professionals. Rather, the betrayal *made* him distrust professionals. Similarly, Cheryl Cole may not resolve to distrust men but rather, after Ashley's unfaithfulness, find herself to be suspicious of them. Jones' view of how victims revise their trust practices to regain metatrust does not account for these kinds of cases.

Local, collateral distrust that is not arrived at through deliberate revision of one's trust practices can be explained by positing an attitude of pessimism on the part of victims towards those they associate with their betrayers. Affective attitudes can be directed towards individuals and institutions that we associate with those whom our attitudes were originally directed towards. For example, after having a wonderful holiday in Scotland I may be disposed to feel positively about any Scottish tourists that I meet on the street in my home town. When I hear their accent I am reminded of the warm hospitality I received while in Edinburgh and I am congenial with the tourists, even though they are not the ones who I engaged with in Scotland. So too, with negative affective attitudes. After having an extremely painful experience at the dentist, I may be negatively disposed towards dentists. When I make a new acquaintance who turns out to be a dentist, I end up regarding her negatively.

I will not attempt to provide a psychological explanation for why affective attitudes can be directed towards those whom we associate with the original object of our attitude. Rather, I will assume that such affective associations occur and explain non-deliberate, local, collateral distrust as resulting from one such association. Like the positive and negative feelings that we have towards those whom we associate with certain things, it is plausible that victims of betrayal can be pessimistic about others who are similar to their betrayers. For example, Cheryl Cole may be pessimistic about other men because the person who betrayed her was a man. When she interacts with men her pessimism resulting from Ashley’s infidelity is triggered and she finds herself being suspicious of other men. She has collateral distrust of them because she associates them with the man who betrayed her. As a result, her interactions with men are affected by her husband’s actions. They do not affect her interactions with all others, in the way that Williams’ murderous actions affect Lloyd’s friend Lisa, but they still significantly affect her ability to freely engage with others.

The damage that betrayal can inflict on its victims’ relationships with various others constitutes damage to their moral agency. Someone with global, collateral distrust will be hesitant about engaging in a full range of interactions with others. She may still be socially active but she will be self-protective, because of her awareness of her vulnerability. In this state an individual is less free to trust others. Someone who experiences local, collateral distrust may be equally self-protective, though only around a select group of others. These limitations can be costly. Not being able to trust others greatly diminishes a victim’s ability to care for whatever matters to her. As Baier points out, humans are by themselves unable to care for all that matters to them (Baier, 1986, p. 231). For example, if Lisa’s extensive, collateral distrust persists she may be reluctant to trust doctors with her health, as well as refraining from trusting law enforcement officers.

To recapitulate, the cases in section 2 show that betrayal does not only damage the reasons its victims have for trusting their betrayers, it can also damage the reasons they have for trusting others more broadly. They may doubt whether others will uphold relevant normative expectations; their own vulnerability to others may feature significantly in their affective experiences; and they may be pessimistic about others that are similar to those who have betrayed them. To the extent that such damage inhibits victims of betrayal from full participation with others, betrayal damages a person's moral agency. And yet, despite these damages, some victims of betrayal come to place trust reasonably again. How is this possible?

Section 4: Recovering Reasonable Trust After Betrayal

As I have shown, betrayal can impede trust in a specific betrayer, or it can impede trust in various others. It can even impede one’s trust of all others. In section 4.1 I explain how it is possible to recover reasonable trust in one’s betrayer. In section 4.2 I explain the recovery of reasonable trust in persons other than one’s betrayer.

Section 4.1: Recovering Reasonable Trust in One's Betrayer

As I explained in section 3.1, distrust can influence the perception of one's reasons for trusting another. Because distrust is self-confirming, if a victim of betrayal distrusts her offender – which victims of betrayal have good reason to do – she is unlikely to perceive any new information about the betrayer as providing her with reason to trust him again. In order for her to place trust reasonably again in her betrayer, the betrayer must show that he is competent and committed to doing what she might count on him to do, but this evidence might be hard to see. In order for the victim to be able to perceive that new information the attitude of pessimism which makes distrust self-confirming needs to be mitigated.

Optimism about trusting an offender can be awakened in, and through, the process of forgiveness. I will suggest that forgiveness is best understood in terms of co-reactive exchange, and show how offender remorse can simultaneously elicit forgiveness and awaken optimism about trusting one's offender.

The "standard approach"[59] to forgiveness in the philosophical literature understands forgiveness as the giving up of resentment for moral reasons.[60] Most proponents of the standard approach understand resentment to be a kind of moral emotion, attitude or mixture

[59] I follow Andrea Westlund in referring to this view as the "standard approach" to forgiveness (Westlund, 2009, p. 507).

[60] Bishop Butler is commonly cited as being one of the first to develop an account of forgiveness as the foreswearing of resentment. (See Murphy (1988, p. 15); Holmgren, (1993, p. 341); Roberts (1995, p. 290); and Hieronymi, (Hieronymi, 2001, p. 529)). But Butler did not argue for what is now the "standard approach." Instead, on his view, forgiveness involves giving up resentment that is out of proportion or exercised without due cause. (Butler, 1827, p. 4) Charles L. Griswold (2007, p. 20), Westlund (2009, p. 509), and Walker (2006, p. 155) note this discrepancy. Nevertheless, the misrepresentation of Butler still has a relatively firm place in the literature.

Though his discussion of forgiveness is brief, Strawson identifies forgiveness as involving the foreswearing of resentment (Strawson, 1974, p. 6). Murphy talks of a "ceasing" to resent in forgiveness (Murphy, 1988, p. 23). McGary shares Murphy's terminology as well (McGary, 1989, p. 349). Calhoun and Hieronymi both say forgiveness involves the *abandonment* of resentment (Calhoun, 1992, p. 87; Hieronymi, 2001, pp. 530-531). And Govier and North both speak in terms of resentment being overcome in forgiveness but while Govier stays within general terms of overcoming, North says that in overcoming resentment we *deny ourselves the right* to resentment (Govier, 2002, p. 43; North, 1987, p. 502). Each of these philosophers uses slightly different terminology but they all take forgiveness to involve some type of "going away" of resentment.

of the two which is roused by offenses done against oneself (Westlund, 2009, p. 507).[61] In developing my account of forgiveness below I adopt Strawson's view of resentment as a participant reactive attitude (Strawson, 1974, p. 4).

On the standard approach it is important that resentment be given up for good reason rather than just fading away. As Hieronymi points out, it would not count as forgiving if one just took a pill that made one's resentment go away (Hieronymi, 2001, p. 530). Good reasons for moving from resentment to forgiveness must not result in condoning, excusing or forgetting the offense that resentment protests, and these reasons must be compatible with victim self-respect, respect for others as moral agents and respect for the rules of morality or the moral order (Murphy, 1988, p. 24). This is because there must be an offense to forgive – so the act cannot be condoned, excused or forgotten – and forgiveness must not undermine the moral context in which one agent violated another.

Most proponents of the standard approach present genuine offender remorse or apology as a good reason for giving up resentment and forgiving.[62] Forgiving because of offender remorse does not excuse, condone or forget the wrongful action. In expressing remorse, the offender recognises that she did something that should not have been done. Forgiving because of offender remorse also acknowledges the moral status of the offender

[61] Strawson provides a seminal discussion of resentment (Strawson, 1974). On his view, resentment is a reactive attitude which responds to expressions of ill will or indifference (Strawson, 1974, p. 10). I will explicate Strawson's understanding of resentment and other reactive attitudes more fully below when developing my account of forgiveness as result of co-reactive exchange. Murphy explains resentment as "a response not to general wrongs but to wrongs against oneself..." (Murphy, 1988, p. 16). McGary does not explicitly say what exactly he takes resentment to be but he implies that it is a moral emotion when he states, "...the person doing the forgiving must undergo an emotional transformation; the person must cease to feel resentful towards the offender in spite of the offender's wrongful action" (McGary, 1989, p. 349). Calhoun follows Strawson in understanding resentment to be a participant reactive attitude (Calhoun, 1992, p. 87). Hieronymi understands resentment to be a kind of moral protest (Hieronymi, 2001, p. 530). I think Govier can be read as implying that resentment is a kind of moral attitude that is part emotion. This can be seen in three claims she makes. "Forgiveness is a matter of working over, amending, and overcoming attitudes..." (Govier, 2002, p. 43). Govier also writes, "Forgiveness requires the overcoming of resentment" (Govier, 2002, p. 50). Lastly, Govier claims, "Resentment is a partial emotion" (Govier, 2002, p. 51). North implies that resentment is a kind of anger or hostile feeling in saying that forgiveness involves overcoming such feelings and then stating that, "If we are to forgive, our resentment is to be overcome..." (North, 1987, p. 502).

[62] The view that resentment is given up because of offender apology is held by Strawson, (1974, p. 6); Hieronymi (2001, p. 548); Murphy (1988, p. 24); and Griswold (2007, pp. 49-50).

and victim, as it expresses that the offender is the kind of person who can offer a morally significant apology and that the victim is the kind of being who can deserve it, receive it and/or elicit it.

Understanding forgiveness as giving up resentment because of offender remorse is not without its challenges, however. Aurel Kolnai has pointed out that while understanding forgiveness as a response to offender remorse avoids conflating forgiveness with condoning wrongdoing, such an understanding risks making forgiveness redundant (Kolnai, 1974, pp. 97-98). Kolnai thinks that in having a change of heart and becoming remorseful an offender ceases to offend and there is therefore no need for forgiveness. As Kolnai puts it,

> Either the wrong is still flourishing, the offence still subsisting: then by 'forgiving' you accept it and thus confirm it and make it worse; or the wrongdoer has suitably annulled and eliminated his offence, and then by harping on it further you would set up a new evil and by 'forgiving' you would only acknowledge the fact that you are no longer its victim. Briefly, forgiveness is either unjustified or pointless. (Kolnai, 1974, pp. 98-99)

For forgiveness to be justified and not collapse into excusing or condoning we cannot change the fact that a wrong has been done and that the victim should not have been wronged. It seems that the only thing that can justify forgiveness – offender remorse – results in forgiveness being redundant since in expressing remorse, the offender ceases to be offensive. That is, if forgiveness comes after remorse, it seems that it arrives too late. This "paradox" of forgiveness (Calhoun, 1992, p. 80; Kolnai, 1974, p. 95) can be resolved.

Forgiveness can be understood as rightly coming after offender remorse without becoming redundant when the resentment, remorse and forgiveness itself are understood as co-reactive attitudes. I will show this by briefly explaining Strawson's view of resentment as a reactive attitude, explicating McGeer's further development of Strawson's view of reactive attitudes as *co-reactive,* and then showing how forgiveness can be a meaningful response to remorse that does not condone wrongdoing when it is understood as the result of co-reactive exchange.

In *Chapter 1* I explained Strawsonian reactivity when introducing the moral vulnerability involved in trust and also when discussing Holton's participant stance approach to trust (Holton, 1994). Here I will focus on the features of Strawson's view, and McGeer's development of it, that are important for my argument in this section.

Strawsonian participant reactive attitudes are attitudes that persons take towards others participating in given relationships with them. They are attitudes which respond to others' states of will towards us as expressed in their attitudes and actions – that is, they track expressions of goodwill, ill will or indifference (Strawson, 1974, pp. 5-10). Resentment is a reactive attitude that persons experience when they take another to have ill will or indifference towards them.

It is plausible that the victims of betrayal in the Smith molestation case and in the Catholic Church case would experience Strawsonian resentment towards their offenders. The offenders in those cases clearly express a lack of goodwill towards their victims.[63] In contrast, on Strawson's view, Lisa would not, strictly, resent her friend's killer. She has no relationship with Williams. Rather, she would presumably have what Strawson identifies as resentment's vicarious analogue – indignation (Strawson, 1974, p. 14). Strawson explains indignation as, "resentment on behalf of another..." (Strawson, 1974, p. 14).

McGeer explicates the attitudes Strawson has identified as not merely backward-looking – that is, they do not only react to another's attitudes – but also forward-looking in that they elicit further responses from the person towards whom they are directed. Insofar as

[63] It may seem strange to claim that the survivor in the Catholic Church case would resent the church that betrayed her, since on Strawson's view, co-reactive attitudes are felt towards persons, but the Catholic Church is an institution, not a person. In *Chapter 3* I showed that it is possible to understand institutions as having character, which makes institutions somewhat similar to individual persons. But understanding institutions as having character is not the same as understanding them to be appropriate recipients of participant reactive attitudes. For them to be capable of participation they need to be able to understand and respond to attitudes that individuals might express to them. It is unclear whether institutions meet this requirement. I leave this concern to one side for now. Before concluding this section I will return to develop an explanation of how it can be appropriate for victims of betrayal to engage in a process of co-reactive forgiveness with institutions – like the Catholic Church.

reactive attitudes elicit reaction they are “co-reactive” (McGeer, 2011, p. 8).

When co-reactive attitudes receive “normatively appropriate reactive response” (McGeer, 2011, p. 9) from others, their function is fulfilled and they are replaced by normatively appropriate reactive responses to the new attitudes that have been received. For example, if you intentionally stomp on my toe, the ill will that your action expresses toward me triggers a reaction of resentment in me. And when I express that resentment to you, my resentment calls for a further response of apology from you, which, if received, elicits a response of forgiveness from me. By reacting, eliciting reaction and responding to that further reaction, co-reactive attitudes are “embedded in dynamic trajectories of reactive exchange” (McGeer, 2011, p. 9).

Understood as a co-reactive attitude forgiveness which responds to offender remorse neither excuses nor condones past wrongdoing. By expressing remorse, the offender acknowledges the wrong, acknowledges that the victim is a morally significant agent whom he should not have wronged, and takes responsibility for that wrong. In his expression of remorse, he does not cease being a wrongdoer but rather expresses something like a desire to change and, I think, a degree of goodwill towards the victim. While that remorse does not “undo” the wrong that was done to the victim, it expresses a change in attitude which can mitigate the victim’s resentment and elicit forgiveness from her. So rather than forgiveness being something which a victim offers to the offender because the remorseful offender now deserves it, co-reactive forgiveness is a reaction to the remorseful offender’s expression of goodwill towards the victim.

Because the betrayed victim’s pessimism influences how she perceives new information about the betrayer, she may not take his expression of remorse to be authentic, at least at first. While the acceptance of offender remorse can elicit forgiveness, the victim’s acceptance of that remorse, her forgiveness of the offender, and the mitigation of her pessimism may be gradual. The victim may need to experience a series of co-reactive exchanges involving offender expression of remorse before she accepts his apology, forgives, and has some optimism about trusting him in the future.

Jones discusses a similar shift from trust to distrust that I think can be extended in understanding this shift from pessimism back to optimism. As I explained in *Chapter 1,* Jones holds that trust can give rise to beliefs which are resistant to evidence (K. Jones, 1996, p. 16). The same is true of distrust; this is its self-confirming nature as I explained in section 3.1 of this chapter. Jones points out that the influence trust and distrust have over one's perception of new information is not limitless (K. Jones, 1996, p. 16).

> If I trust you, I will, for example, believe that you are innocent of the hideous crime with which you are charged, and I will suppose that the apparently mounting evidence of your guilt can be explained in some way compatible with your innocence. [...] Given enough evidence, my trust can be shaken and I can come to believe that you are guilty. (K. Jones, 1996, p. 16)

Similarly, after a series of co-reactive exchanges with one's offender, wherein a victim gathers evidence that the offender truly is remorseful, a victim may come to accept the offender's expression of remorse as authentic. When that happens, her resentment can be traded for forgiveness, her pessimism can be mitigated, and optimism about trusting the offender again may begin to be awakened. Authentic offender remorse can do this work because it shows that the offender can take responsibility for his moral interactions; that he cares about upholding the normative expectations that he is obliged to uphold in moral relationships; and that he has a degree of goodwill towards the victim. It is through a series of co-reactive exchanges, then, that victims can come to forgive, have their pessimism moderated, and be in a position where they can perceive new information about an offender that might give them reason to trust him. This does not mean that the victim will necessarily trust the offender again. She may have some optimism about trusting him but still resist accepting vulnerability to him because she still thinks he is relevantly incompetent or uncommitted to doing what she would otherwise count on him to do. But by coming to forgive the offender, the attitude that perpetuates her distrust can be mitigated so that she can at least perceive the offender without the filter of distrust.

My view of recovering reasonable trust in one's betrayer requires co-reactive exchange between persons – specifically, betrayers and those who have been betrayed. However, as the Catholic Church case I discussed in section 2 shows, not all betrayers are persons. The survivor in that case was betrayed both by the priest who abused her and by the Catholic Church which enabled that abuse. Since institutions are not persons it might be thought that they cannot be participants in co-reactive exchange. However persons and institutions do seem to engage in co-reactive exchange. Individuals sometimes demand remorse in the form of institutional recognition of, and apology for, the harm inflicted in institutional contexts. Apologies are delivered. For twenty years individuals involved in the Child Migrants Trust have campaigned for recognition of the harm done to child migrants by government institutions (Constantine, 2009). Sometimes, institutions respond positively to the call for recognition. As discussed in *Chapter 3,* in 2009 Prime Minister Kevin Rudd made a formal apology on behalf of the Federal Government of Australia to persons abused in State governed "homes" (Rudd, 2009). So it seems that co-reactive exchange can, and does, involve institutions. An explanation is required then for how institutions can engage in co-reactive exchange.

With some interpretation, the understanding of co-reactive exchange which I have adopted from McGeer can be applied to institutions. Co-reactive attitudes are "well-targeted" (McGeer, 2011, p. 7) if their recipients can understand the message of the attitudes being expressed to them and respond in ways that show "normative awareness of demands being made of them" (McGeer, 2011, p. 7). The challenge presented above is precisely that institutions qua organizations cannot understand the message of our reactive attitudes and cannot respond in ways showing normative awareness about the content expressed by our attitudes.

My response to this challenge is that the role-holders representing institutions are capable of understanding our message and responding in normatively appropriate ways. Further, individuals usually address their attitudes about institutions to such role-holders. For example, when we are resentful or indignant at a government for the way it is handling national health care we may address our concerns to the relevant government representative such as the minister and urge her to reform policies and legislation. Role-holders who recognise others' legitimate co-reactive attitudes can work to have the offending institutions respond in a way that shows awareness of the moral demands being made. They may make formal apologies on behalf of the institutions they represent. Or they may work to reform institutional policies and practices so that individuals are treated with appropriate moral regard. Then, after institutional remorse has been expressed – through apology and reform – as in relations between victims and individuals who betrayed them, victims betrayed by institutions may respond with forgiveness and become more optimistic about trusting the institution again. On my view, then, institutions can engage in co-reactive exchange insofar as their role-holders receive the messages expressed by reactive attitudes such as resentment and respond appropriately.

The offender remorse that aids victims in recovering reasonable trust of their betrayers is not always available, however. An offender may be unavailable for a victim to engage in co-reactive exchange with him. He may be dead, incarcerated or otherwise out of touch. Or if the offender is not out of touch, it may be inappropriate or too emotionally painful for the victim to interact with him. For example, given the manipulation and abuse she endured, the survivor in the Catholic Church case may experience debilitating anger or fear when she considers engaging in co-reactive exchange with the priest who betrayed her.

Or she may find that when she does go to express her resentment to him, the manipulative power that he used to have over her begins to come back and it is difficult for her to avoid re-enacting her old role as someone who could not stand up to him. Further, the offending betrayer may be unapologetic and beyond remorse. The priest may be so dysfunctional and deluded that he will not believe that his actions were wrong. For these reasons victims of betrayal may not receive the offender remorse that they need to recover reasonable trust in those who betrayed them.

When offender remorse is unavailable, co-reactive resolution between the offender and the victim cannot be reached. That can be a deeply disturbing outcome for the victim. It can be hard to live with the strong anger and devastating emotional pain caused by someone who is unremorseful or unavailable for co-reactive resolution. But as I will show in the next section, it is possible for victims of betrayal to recover reasonable trust in persons *other than* those who betrayed them. Despite the victim's relationship with their offender being beyond repair, victims of betrayal may still be able to flourish in other relationships.

Section 4.2: Recovering Reasonable Trust in Those Other Than One's Betrayer

It is possible for victims of betrayal to recover reasonable trust in people other than their betrayers. I begin by explaining how victims can recover reasonable trust toward those in whom they have local, collateral distrust. Then I explain how those with global, collateral distrust can trust others reasonably again.

As I explained in section 3.2, local collateral distrust can come about in two ways. It can result from a victim of betrayal deliberately revising their trust practices so that they no longer trust a certain kind of individual. Or it can occur because victims of betrayal are pessimistic about those whom they associate with their betrayers. The recovery of reasonable trust in both situations needs to be explained.

A person might amend her resolve to withhold trust from a certain group of others when she recognizes some members of that group as being relevantly competent and committed to doing what she would count on them to do were she to trust them. I will argue to this conclusion by considering what it would take for the survivor in the Catholic Church case to come to trust some clergy again.

I think it is plausible that the survivor could trust some clergy if she consistently observed them as practicing what they preach and caring appropriately for their parishioners. For example, the survivor sought help from the bishop in charge of her diocese. He provided some financial assistance to aid her with the costs of the counselling she was receiving (Fleming, 2007, p. 159). The survivor tells that she later discovered that the bishop was, himself, a "sexual predator" (Fleming, 2007, p. 164). Receiving that knowledge about the bishop would have strengthened the survivor's distrust of all clergy. In contrast, if the bishop had proven not to be an offending abuser himself and, instead, worked to hold abusive priests to account, the survivor might have come to trust him. If she had a number of positive experiences with other clergy, she might come to see that not all clergy are abusive and that some can be trusted. With that knowledge, it is plausible that she could amend her revised trust practices so that she does not distrust all clergy but rather considers them on a case by case basis. A similar explanation can be given about the recovery of reasonable trust when local, collateral distrust is based on an attitude of pessimism about those associated with one's betrayer.

Pessimism about those associated with one's betrayer can be reduced when a victim receives *enough evidence* that is contrary to her collateral distrust. I will return to explicate what will count as enough evidence after arguing that *some* amount of evidence is able to overturn one's pessimism. In section 4.1 I adapted an argument from Jones about how quasi-perceptual trust can be shaken, given enough evidence (K. Jones, 1996, p. 16). I explained that, with enough evidence, the pessimism involved in distrust can also be shaken. This argument can also be used to explain the possibility of recovering reasonable trust in persons whom one negatively associates with one's betrayer. If a victim of betrayal receives a sufficient amount of evidence that those towards whom she has local, collateral distrust are, in fact, relevantly competent and committed to doing what she might count on them to do, her pessimism about them can be mitigated. Consider Cheryl Cole's presumed distrust of men. If she observes some men to be consistently competent and committed to their romantic partners, her generalised pessimism about all men may recede. This need not mean that she will automatically trust any man. It just means that she may cease distrusting all of them and be free to consider whether individual men might be competent and committed to do what she might count on them to do.

What will count as *enough* evidence for a victim to cease being pessimistic about those toward whom she has local, collateral distrust will vary. If a person has been betrayed multiple times by members of a certain group, she is likely to require more evidence that others are relevantly competent and committed before her pessimism will recede. For example, if Cheryl Cole had been cheated on by multiple men throughout her life she would require much more evidence that men can be relevantly competent and committed in romantic relationships than if she was only cheated on by her husband, Ashley. She may still need a significant amount of evidence in order to trust other men after being betrayed by Ashley. But she would need more evidence if her relational history were different.

Victims of betrayal may never accrue the evidence needed for them to recover reasonable trust in those towards whom they have local, collateral distrust. Their revised trust practices may remain unamended. And their pessimism about those they associate with their betrayers may persist. It is at least possible, though, for them to receive the evidence necessary for them to recover that trust in others. However, when a victim's default trust and basal security have been damaged, more will be needed before they can recover reasonable trust in others.

For victims who question their own judgement and have had their default trust and basal security damaged, recovering reasons to trust again may seem close to impossible. They need to regain confidence in their judgements, have reason to think that others will uphold relevant normative expectations, and have some reason to risk being betrayed again despite being very aware of their own vulnerability. In short, they need to regain confidence as a truster.

Victims of betrayal can have their confidence as trusters scaffolded by the moral community around them. By “the moral community” I mean those individuals, institutions and institutional role-holders with whom a person engages in moral interaction. There are three main ways in which members of the moral community can scaffold one’s confidence as a truster: they can validate that what was done to the victim was, in fact, wrong; they can uphold relevant normative expectations themselves; and they can act as models for judging who it is reasonable to trust.

First, individuals and institutions can help victims to recover reasonable trust in others by validating that what was done to them should not have been done. Walker has pointed to this mechanism for recovery: “Third parties play a crucial role in signalling to those wronged, to wrongdoers, and to each other that an action requires a response that reasserts the norms and recognizes victims and wrongdoers as such” (Walker, 2006, p. 95). Individual members of the moral community can do this important “signalling” work by expressing indignation at what has happened to the victim and calling for betrayers to recognize the wrongs they have done. For example, Cheryl Cole’s friends may gather around her, express their anger at what her husband did, and confront him about his treatment of her.

Institutions can also acknowledge that what was done to a victim was wrong and should not have happened. They can do this by issuing statements supporting victims of betrayal and condemning their betrayers. For example, as I stated in section 4.1 the Child Migrants Trust has campaigned for recognition of the harm done to child migrants by government institutions (Constantine, 2009).

Government institutions themselves can support victims of betrayal by acknowledging that the actions done to the victims were wrong and should not have been done. Prime Minister Rudd's apology to persons abused in State governed "homes" is a case in point. As a chief representative of the Australian Federal Government, Mr. Rudd acknowledged that children entrusted to that government's care were treated wrongly and that the government was partially responsible for the offenses that were committed (Rudd, 2009). That acknowledgement was, arguably, an attempt to scaffold the recovery of persons who were negatively affected by State governed "homes."

Second, the moral community can scaffold a victim's confidence as a truster by upholding normative expectations themselves. While being self-protective, a victim of betrayal who has global, collateral distrust may observe that individuals and institutions around her do, in fact, act as they ought to. Given this observation, she may regain some confidence about trusting others. For example, Lloyd's friend Lisa may notice that, while every individual is a potential betrayer, few of them actually do betray others. That realization may be one of the first steps in her recovering reasonable trust of others.

Similarly, persons who have been betrayed by institutions may regain confidence that some institutions will uphold relevant normative expectations by observing that institutions other than the one that betrayed them do in fact act well with regard to individuals counting on them. For example, the survivor in the Catholic Church case may observe that the role-holders in a different religious institution do not abuse members of their congregation and so come to regain confidence in trusting other religious institutions.

Third, a victim's confidence as a truster can be regained when their confidence as a judge of trustability is strengthened. This can be done through re-learning who it is reasonable to trust by watching the trust practices of others; that is, by observing whom others trust and whom they distrust. This method for regaining confidence in one’s judgements seems to face a dilemma. It seems that the victim of betrayal must trust those on whom she models her judgements. But if she has global, collateral distrust, she does not trust anyone. Thus it seems she must first be able to trust in order to use this method to regain her confidence as a truster. However, this dilemma can be solved. The victim can at first merely rely on those whose trust practices she uses to model her judgments, but then if those judgments are repeatedly confirmed she may gradually regain her self-confidence as a judge of trustability and so may come to trust others again. For example, recall Lisa’s testimony that there was no one left who she could trust. Lisa may see her mother trusting some armed services personnel and say to her, “How can you trust them after what happened to Jessica?” If Lisa takes her mother’s answer to be well grounded and not mistaken or mischievous, she will have at least some reason to rely on her mother’s judgements of trustability in the future. Over time, Lisa can then re-learn how to judge who it is reasonably safe to trust by watching her mother. When her judgements of trustability cohere with what she takes to be her mother’s reliable judgements, Lisa has some reason to think that she is a good judge of trustability. All of this has taken place without Lisa having to step out and actually risk trusting another person.

In this section I have explained how it is possible for victims of betrayal to recover reasonable trust in those other than their betrayers. If they have enough evidence contrary to their local, collateral distrust, they can amend relevant revisions they had previously made to their trust practices. With that evidence, their pessimism towards those associated with their betrayers can be overcome. Victims who have had their confidence as a judge of trustability, default trust, and basal security damaged can also recover reasonable trust in others. Their trust can be scaffolded when the moral community affirms that what was done to them was wrong, upholds normative expectations and provides models of how to judge who it is reasonable to trust. All of this recovery can be done without victims having recovered reasonable trust in their betrayers.

My explanation of how it is possible to recover reasonable trust after betrayal shows that recovery cannot be done on one’s own. To recover trust in one’s betrayer, both victim and betrayer must engage in a series of co-reactive exchanges. For victims to trust those towards whom they have local, collateral distrust they must at least be able to observe the competence and commitment of those they currently distrust. Victims need not necessarily directly engage with those they distrust, but they will not be able to gain evidence that those they distrust are, in fact, competent and committed in relevant ways unless they are able to observe how those persons act in relations with others. And for a person to recover from global, collateral distrust the moral community must provide the scaffolding to enable her to learn how to reasonably trust again. In order to recover reasonable trust, we need each other.

Conditions for recovering reasonable trust after betrayal will not always be in place. As I have already pointed out, not all betrayers will take responsibility for their actions and express authentic remorse to their victims. Further, not all communities will aid in scaffolding the recovery of those who have been betrayed. The individuals and institutions surrounding victims may fail to acknowledge the wrongs done to the victims, and they may fail to uphold normative expectations themselves. But even if conditions for recovering reasonable trust *are* in place, some victims of betrayal may have had their ability to trust so severely damaged that it is essentially irrecoverable. Despite receiving their offender’s expression of authentic remorse and despite their community’s scaffolding work, such a victim may continue to have pervasive distrust of others. The damage which betrayal inflicts can be lasting. However, as I have shown, in some cases at least, reasonable trust after betrayal can be recovered.

Conclusion

The concept of betrayal has to date been largely overlooked in the philosophical literature on trust. This, I have argued, is problematic, since notions of trust themselves turn on assumptions about betrayal. This thesis has been motivated in part by a need to rectify this deficit in the literature. As I have shown, a clear account of betrayal is required in order to explain the distinction between trust and mere reliance, to understand the damage that betrayal can inflict on trust, and to explain the possibility of recovering reasonable trust after betrayal.

In *Chapter 1* I set aside the concept of betrayal in order to provide an analysis of trust. I showed that a satisfactory understanding of trust needs to account for a variety of phenomena, including the difference between trust and mere reliance. After showing that four influential approaches to understanding trust cannot account for some trust-relevant phenomena, I suggested that we need to reconsider what kind of concept trust is. Rather than treating trust as a concept that can be reduced to a set of necessary and sufficient conditions, I developed an account of trust as a cluster concept. I argued that this approach to understanding trust can account for the variety in trust phenomena and for the difference between trust and mere reliance. The central implications of this section of my argument are that we should change the way we think about trust – about what it is and how it works. Not all instances of trust will have the same features, and it will not always be clear whether an interaction should be understood in terms of trust or mere reliance. Border-line cases exist, and understanding trust as a cluster concept can accommodate these messy cases.

In *Chapter 2* I argued that the trustworthy are not the only ones it is reasonable to trust. I arrived at this conclusion in three steps. First, I considered accounts of trustworthiness from McLeod (2011); Hardin (1991, 1996, 2002); Baier (1986, 1994); Jones (2011); and

Pettit (1995), which understand trustworthiness as a matter of competence and commitment to do what another is counting on you to do. Second, after showing that persons can be relevantly competent and committed without being trustworthy, I adapted Potter’s (2002) character-based understanding of trustworthiness. Third, I showed that it can be reasonable to trust someone without knowing much about their character. As such, the trustworthy are not the only ones it is reasonable to trust. I developed the concept of trustability to identify those whom it is reasonable to trust and I identified trustworthiness as one kind of trustability. My argument in *Chapter 2* shows that there is not an exclusive relation between reasonable trust and trustworthiness. Rather, being trustworthy is a specific way of being one whom others have reason to trust.

In *Chapter 3* I dealt with three challenges that complicate the application of the concepts of trust, trustability and trustworthiness to institutional contexts. I used the Catholic Church example to show that relations involving individuals, institutions and institutional role-holders *can* be sufficiently similar to paradigm cases of trust to be instances of trust themselves. Then, contrary to Hardin’s (2002) epistemic limitations claim, I showed that it *is* possible to have knowledge that is sufficient to trust institutions reasonably. Using O’Neill’s concepts of “intelligent accountability” and “active checking”, I identified mechanisms that can encourage institutional role-holders to be competent and committed to doing what others are counting on them to do (O'Neill, 2002b, pp. 58, 77). When policies which establish such accountability and enable such checking are created and enacted, individuals can have enough information to place trust reasonably in institutions and their role-holders. Having shown that the concepts of trust and trustability apply to institutional contexts, I demonstrated how the character-based account of trustworthiness that I developed in *Chapter 2* is applicable to institutions. Adapting T. Jones’ (T. Jones, 2007) functional understanding of groups, I constructed a functional account of institutional character. On my view, institutional character consists of an institution's function, the way it fulfils its functions, and the roles and policies disposing it to fulfil those functions in the ways that it does. I then used Oakley and Cocking's (1998) view of virtuous professional roles in conjunction with the character-based view of trustworthiness that I adapted from Potter (2002) in *Chapter 2* to develop an account of institutional integrated, virtuous character. My analysis in *Chapter 3* shows that trust-relevant concepts which are often associated with interpersonal interaction are salient in institutional contexts as well.

In *Chapter 4* I provided a substantive account of the concept of betrayal. After analysing betrayal as a disappointment of normative, "constitutive", expectations that shape particular relational domains, I identified similarities between that analysis and Keller's (2007) understanding of disloyalty. In light of the similarities and differences between betrayal and mere disloyalty, I argued that betrayal is best understood as a *kind* of disloyalty which involves a failure to uphold normative expectations in a way that undermines special relationships. Having developed an account of betrayal, I reconsidered the claim that vulnerability to betrayal is distinctive of trust. I showed that it is possible to be vulnerable to betrayal when one merely relies on another. Betrayal, then, should not be exclusively associated with trust. However I also explained why it is understandable to make that association: trust, but not mere reliance, can establish the special relationships which betrayal undermines.

In *Chapter 5* I showed that betrayal can result in direct distrust; local, collateral distrust; global, collateral distrust; loss of confidence as a judge of trustability; and negative emotional responses. Given these effects I argued that betrayal damages the reasons its victims had for trusting their betrayers and the reasons they had for trusting others more broadly. Betrayal gives its victims reason to judge their offenders to be untrustable. It influences the way they perceive new information about their betrayers. It can damage their perception of persons whom they associate with their betrayers. It can reduce its victims' confidence as judges of trustability. And it can strike at their default trust, basal security, and metatrust. As a result of the aforementioned damage, I discussed how it is possible for victims of betrayal to recover reasonable trust – both in their betrayers and in others more broadly.

In explaining how victims can come to trust their betrayers again, I argued that forgiveness is best understood as a co-reactive attitude which responds to offender remorse. This understanding of forgiveness solves the forgiveness "paradox" (Calhoun, 1992, p. 80; Kolnai, 1974, p. 95). As a co-reactive attitude, forgiveness neither excuses nor condones wrongdoing and it is not something that is earned by offenders. Rather, it is a reaction that occurs because the threat which one's resentment protested is no longer present.

In explaining the possibility of recovering reasonable trust in individuals and institutions other than one's offender, I showed that members of the moral community can play a significant, scaffolding role in that recovery. They can affirm that what was done to the victim was, in fact, wrong and should not have been done. They can show the victim that not everyone fails to uphold normative expectations. And they can model good judgements of who it is reasonable to trust.

My arguments in *Chapter 5* show that, while it is possible to recover reasonable trust after being betrayed, that recovery cannot be done on one's own. Coming to place trust reasonably when one has been betrayed requires involvement from offending betrayers and/or the moral community.

Given the account of betrayal that I have provided in this thesis, that concept can be used to do more, substantive, work. Briefly, here are just two ways my work could be taken further. First, beyond being able to test claims about the distinctness of trust and mere reliance, we now have a basis for considering whether talk about "feeling betrayed" should be taken at face value. For example, Walker argues that default trust is, in fact, a kind of trust because persons feel indignation, resentment and *betrayal* when it is disappointed (Walker, 2006, p. 85). Walker does not, however, explain the concept of betrayal or discuss whether feeling betrayed means that one really has been betrayed. Arguably, providing a substantive account of betrayal better positions us to assess claims about “feeling betrayed.” We can test whether those who feel betrayed really have been betrayed. And we can consider whether such feelings identify violations of trust or not.

Second, Walker argues that restorative justice exemplifies repair of damaged moral relations: “It captures the truth at the centre of moral repair: morality as a practice of human life is embodied in relationships that require confidence, trust, hope, and in interactions that in turn renew these morally sustaining attitudes” (Walker, 2006, p. 229). My account arguably complements Walker’s understanding of the role that restorative justice can play in recovering trust in victims who have been wronged and betrayed, including by institutions and their role-holders. On my account betrayal can be understood as a violation of normative expectations. The implication of this account is that restorative justice works to restore trust, in part by acknowledging and recognizing that the victim’s and moral community’s legitimate normative expectations have been violated and by reaffirming the moral community’s commitment to upholding those norms.

I began the introduction to this thesis by noting the importance of trust to living a flourishing life. My analysis of trust and trustworthiness explains why that is the case. My analysis of betrayal, and of the interplay between trust and betrayal, also explains why betrayal, and the damage it is capable of causing, can be so undermining of the trust in others that is essential for human flourishing.

Bibliography

Austen, J. (1992). *Sense and Sensibility*. Hertfordshire, UK: Wordsworth Editions.

Baier, A. (1986). Trust and Antitrust. *Ethics, 96* (2), 231-260.

---- (1992). Trusting People. *Philosophical Perspectives, 6*, 137-153.

---- (1994). *Moral Prejudices*. Cambridge, Massachusetts: Harvard University Press.

Boshoff, A. a. S. N. (2010). With a Tweet, It's Over for Cheryl and Ashley Cole: They Were So in Love With Themselves, The Marriage Never Stood A Chance. Retrieved 13.08.2011, 2011, from http://www.dailymail.co.uk/tvshowbiz/article-1281759/Ashley-Cheryl-Cole-divorce-The-marriage-stood-chance.html

Brown, G. (2010). Statement: Child Migration Gordon Brown. Retrieved 09.08.2011, 2011, fromhttp://www.parliamentlive.tv/Main/Player.aspx?meetingId=5831&st=12%3A33%3A55

Butler, B. J. (1827). Fifteen Sermons Preached at the Rolls Chapel. In L. Dagg (Eds.) Available from http://anglicanhistory.org/butler/rolls/08.html

Calhoun, C. (1992). Changing One's Heart. *Ethics, 103* (October, 1992), 76-96.

The Cambridge Spy Ring. (1999). Retrieved 17.05.10, 2010, from http://news.bbc.co.uk/2/hi/special_report/1999/09/99/britain_betrayed/444058.stm

Cocking, D., & Kennett, J. (1998). Friendship and the Self. *Ethics, 108* (April 1998), 502-527.

COHRE. (2010). UN Human Rights Council Adopts Resolution on Housing Rights and 'Mega-Events'. Retrieved 10.08.11, 2011, from http://www.cohre.org/news/press-releases/un-human-rights-council-adopts-resolution-on-housing-rights-and-mega-events

Constantine, S. (2009). UK Child MIgrants Apology Planned. Retrieved 24.09.11, 2011, from http://news.bbc.co.uk/2/hi/uk_news/politics/8361025.stm

de Sousa, R. (1987). *The Rationality of Emotion*. Cambridge, Massachusetts: The MIT Press.

---- (2010). "Emotion", The Stanford Encyclopaedia of Philosophy, from <http://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=emotion>

Donnersmarck, F. H. v. (Writer). (2006). The Lives of Others (Das Leben der Anderen. Germany.

Dutter, B. (2001). British Child Migrants Get Apology for Abuse. Retrieved 09.08.2011, 2011, from http://www.telegraph.co.uk/news/worldnews/australiaandthepacific/australia/1327634/British-child-migrants-get-apology-for-abuse.html

Engels, F. (2008). *The Condition of the Working-Class in England in 1844* (F. K. Wischnewetzky, Trans.). New York: Cosimo.

Fleming, P., Sue Lauber-Flaming and Mark T. Matousek. (2007). *Broken Trust*. New York: The Crossroad Publishing Company.

Funder, A. (2002). *Stasiland*. Melbourne, VIC. Australia: The Text Publishing Company.

Gambetta, D. (1988). Can We Trust Trust? . In D. Gambetta (Ed.), *Trust: making and breaking Cooperative Relations*. New York: Blackwell.

Gill, A. (1997). *Orphans of the Empire*. Alexandria, NSW: Millennium Books.

Govier, T. (1992a). Distrust as a Practical Problem. *Journal of Social Philosophy, 23* (1), 52-63.

---- (1992b). Trust, Distrust, and Feminist Theory. *Hypatia, 7* (1), 16-33.

---- (1994). Is It a Jungle Out There? Trust, Distrust and the Construction of Social Reality. *Dialogue: Canadian Philosophical Review/Revue canadienne de philosophie, 33* (02), 237-252.

---- (2002). *Forgiveness and Revenge*. New York: Routledge.

Griswold, C. L. (2007). *Forgiveness: A Philosophical Exploration*. New York: Cambridge University Press.

Hardin, R. (1991). Trusting Persons, Trusting Institutions. In R. J. Zeckhauser (Ed.), *Strategy and Choice*. Cambridge, Massachusetts: The MIT Press.

---- (1996). Trustworthiness. *Ethics, 107* (1), 26-42.

---- (2002). *Trust and Trustworthiness* (Vol. IV). New York: Russell Sage Foundation.

---- (2006). *Trust*. Cambridge, UK: Polity Press.

---- (Ed.). (2004). *Distrust* (Vol. VIII). New York: Russell Sage Foundation.

Hieronymi, P. (2001). Articulating an Uncompromising Forgiveness. *Philosophy and Phenomenological Research, 62* (3), 529-555.

---- (2008). The Reasons of Trust. *Australasian Journal of Philosophy, 86* (2), 213-236.

Holmgren, M. R. (1993). Forgiveness and the Intrinsic Value of Persons. *American Philosophical Quarterly, 30* (4), 341-352.

Holton, R. (1994). Deciding to Trust, Coming to Believe. *Australasian Journal of Philosophy, 72* (1), 63-76.

Horsburgh, H. J. N. (1960). The Ethics of Trust. *The Philosophical Quarterly, 10* (41), 343-354.

Jackson, R. L. (2000). The Sense and Sensiblity of Betrayal: Discovering the Meaning of Treachery through Jane Austen. *Humanitas, XIII* (2), 72-89.

Jones, K. (1996). Trust as an Affective Attitude. *Ethics, 107* (1), 4-25.

---- (1999). Second-hand Moral Knowledge. *The Journal of Philosophy, 96* (2), 55-78.

---- (2004). Trust and Terror. In P. DesAutels & M. U. Walker (Eds.), *Moral Psychology: Feminist Ethics and Social Theory*: Rowman & Littlefield.

---- (2011). Trustworthiness. *Submitted but as yet unpublished with permission 13 July, 2011*.

Jones, T. (2007). Numerous Ways to be an Open-Minded Organization: A Reply to Lahroodi. *Social Epistemology, 21* (4), 439-448.

Karan, P. (2010). World War 1: The Christmas Truce of 1914. Retrieved 12.01.11, 2011, from http://www.suite101.com/content/world-war-1-the-christmas-truce-of-1914-a322640

Keller, S. (2007). *The Limits of Loyalty*. New York: Cambridge Univ. Press.

Kolnai, A. (1974). Forgiveness. *Proceedings of the Aristotelian Society, 74* (1973 - 1974), 91-106.

Lahroodi, R. (2007). Collectiv Epistemic Virtues. *Social Epistemology, 21* (3), 281-297.

Larson, D. W. (2004). Distrust: Prudent, If Not Always Wise. In R. Hardin (Ed.), *Distrust*. New York: Russell Sage Foundation.

Lewis, D. (1997). Finkish Dispositions. *The Philosophical Quarterly, 47* (187), 143-158.

Loach, J. (Writer). (2010). Oranges and Sunshine [Film].

Lowe, A. (2010). Molested Students Reveal Betrayal. Retrieved 26.11.10, 2010, from http://www.theage.com.au/victoria/molested-students-reveal-betrayal-20100429-twlt.html

McCahill, W. (2009). Hot Rumor: Tiger's Cheating. Retrieved 17.05.10, 2010, from http://www.newser.com/story/74848/hot-rumor-tigers-cheating.html

McGary, H. (1989). Forgiveness. *American Philosophical Quarterly, 26* (4), 343-351.

McGeer, V. (2002). Developing Trust. *Philosophical Explorations, V* (I), 21-38.

---- (2008). Trust, Hope and Empowerment. *Australasian Journal of Philosophy, 86* (2), 237-254.

---- (2011). Co-Reactive Attitudes and the Making of Moral Community. In R. a. C. M. Langdon (Ed.), *Emotions, Imagination and Moral Reasoning*. New York: Psychology Press.

McLeod, C. (2011). "Trust", The Stanford Encyclopedia of Philosophy. Spring 2011 Edition. Retrieved 20.06.2011, from http://plato.stanford.edu/archives/spr2011/entries/trust/

Mullin, A. (2005). Trust, Social Norms, and Motherhood. *Journal of Social Philosophy, 36* (3), 316-330.

Murphy, J. (1988). *Forgiveness and mercy*. Sydney: Cambridge University Press.

Nguyen, L. a. M. H. (2010). Pain, hatred, betrayal and disgust — victims tell Williams of evil he did. Retrieved 26.11.10, 2010, from http://www.ottawacitizen.com/story_print.html?id=3702523&sponsor

North, J. (1987). Wrongdoing and Forgiveness. *Philosophy, 62* (242), 449-508.

O'Neill, O. (2002a). *Autonomy and Trust in Bioethics*. New York: Cambridge University Press.

---- (2002b). *A Question of Trust*. New York: Cambridge University Press.

Oakley, J., & Cocking, D. (2001). *Virtue Ethics and Professional Roles*. New York: Cambridge Univ. Press.

Penglase, J. (2007). *Orphans of the Living*. Fremantle, WA. Australia: Fremantle Press.

Pettit, P. (1995). The Cunning of Trust. *Philosophy and Public Affairs, 24* (3), 202-225.

Potter, N. (2002). *How Can I Be Trusted?: A Virtue Theory of Trustworthiness*. Oxford, England: Rowman and Littlefield.

Rees, S. (2009). The Christmas Truce. Retrieved 12.01.11, 2011, from http://www.firstworldwar.com/features/christmastruce.htm

Richards, F. (2006). Christmas in the Trenches, 1914. Retrieved 12.01.2011, 2011, from http://www.eyewitnesstohistory.com/trenches.htm

Roberts, R. C. (1995). Forgivingness *American Philosophical Quarterly, 32* (4), 289-306.

Rudd, K. (2009). Transcript of Kevin Rudd's Apology to Forgotten Australians. Retrieved 17.11.09, 2009, from http://www.heraldsun.com.au/news/national/transcript-of-kevin-rudds-apology-to-forgotten-australians/story-e6frf7l6-1225798255277

Stoljar, N. (1995). Essence, Identity, and the Concept of Woman. *Philosophical Topics, 23* (2).

Stout, R. (2005). *Action*. Chesham, England. : Acumen.

Strawson, P. F. (1974). *Freedom and Resentment and Other Essays*. London: Methuen.

Thomas, L. (1987). Friendship. *Synthese, 72*, 217-236.

Topping, A. (2010). Cheryl COle Leaves Husband Ashley. Retrieved 23.07.10, 2010, from http://www.guardian.co.uk/culture/2010/feb/23/cheryl-cole-leaves-ashley

Ullmann-Margalit, E. (2004). Trust, Distrust, and In Between. In R. Hardin (Ed.), *Distrust*. New York: Russell Sage Foundation.

Walker, M. (2006). *Moral Repair: Reconstructing Moral Relations after Wrongdoing*. New York: Cambridge University Press.

Westlund, A. (2009). Anger, Faith and Forgiveness. *The Monist, 92* (4), 507-536.

Wittgenstein, L. (1958). *Philosophical Investigations*. Oxford: Blackwell.

Wolf, S. (1980). Asymmetrical Freedom. *The Journal of Philosophy, 77* (3), 151-166.

Woods, T. (2010). Text of Tiger Woods' Public Apology. Retrieved 17.05.10, 2010, from http://www.abc.net.au/sport/stories/2010/02/20/2825395.htm

Wright, S. (2009). Trust and Trustworthiness. *Philosophia, 38* (3), 615-627.