WHAT IS TRUSTWORTHINESS?

Chris Kelp & Mona Simion
Cogito Epistemology Research Centre, University of Glasgow

Forthcoming in *Nous*

Abstract. This paper develops a novel, bifocal account of trustworthiness according to which both trustworthiness *simpliciter* and trustworthiness to *phi* are analysed in terms of dispositions to fulfil one's obligations. In the case of trustworthiness to *phi*, the relevant obligations are one's obligations to *phi*. In the case of trustworthiness *simpliciter*, the relevant obligations are one's contextually-salient obligations. We also offer a systematic account of the relation between the two types of trustworthiness, and the first accounts of degrees of trustworthiness and permissible trustworthiness attribution in the literature to date.

1 Introduction

What does it take to be trustworthy? At a first glance, trustworthiness seems like a complicated affair. First, trustworthiness seems complicated for the person who aims to be trustworthy in virtue of the fact that it is, intuitively, hard to achieve: it's a property that isn't easy to come by. Second, it seems like a complicated topic for the theorist to shed light on, because there seem to be many ways and domains in which one can be more or less trustworthy. If so, the function from these diverse ways of being trustworthy to trustworthiness *simpliciter* seems like a hard one to figure out.

The central aim of this paper is to develop a novel account of trustworthiness. The key idea is, very roughly, that trustworthiness is a disposition to fulfil one's obligations.

Besides offering a novel account of trustworthiness, we also develop a novel way of approaching the phenomenon, on two counts. First, extant accounts in the literature have focused either on what we will call trustworthiness *simpliciter*, i.e. on what it takes for it to be true e.g. of Ann that Ann is trustworthy; or else they have focused on what we will call trustworthiness to *phi*, i.e. what it takes for it to be true e.g. of George that George is trustworthy when it comes to washing the dishes. However, the relation between the two has remained underexplored. Our account remedies this defect. It offers a systematic account of the relation between the two.

Second, extant accounts of trustworthiness have mostly focused on what we will call *outright* trustworthiness, i.e. what it takes for it to be true of e.g. Ann that Ann is trustworthy, or trustworthy to phi. They have spent little time on a systematic account of *degrees* of trustworthiness, i.e. what it takes for it to be true of e.g. Ann and George that Ann is more trustworthy than George. Once again, our account supplies this lack and offers a systematic account of degrees of trustworthiness.

Here is the game plan for this paper. Section 2 takes a look at extant accounts of trustworthiness. Our central aim here is to identify a trilemma for accounts of trustworthiness: they threaten to be either too demanding, or too permissive, or counterintuitively disjoint, in that they fail to account for the plausible relation between different types of trustworthiness. Section 3 develops our novel account of trustworthiness. Section 4 shows how our account points the way towards an attractive solution to the trilemma. We also compare our account with the competition. More specifically, we will argue that our account can not only avoid the problems of rival views but also accommodate their most important insights.

2 Trustworthiness and Demandingness

Trust can be a two-place or a three-place relation. In the former case, it is a relation between a trustor and a trustee, as in "Ann trusts George." Two-place trust seems to be a fairly highbrow affair: when we say that Ann trusts George, we seem to attribute a fairly robust attitude to Ann, whereby she trusts him in (at least) several respects. In contrast, three-place trust is a relation between a trustor, a trustee, and something entrusted as in "Ann trusts George to wash the dishes". Three-place trust is a less involved affair: when we say that Ann trusts George to wash the dishes, for instance, we need not say much about their relationship otherwise.

This contrast is preserved when we switch from focusing on the trustor's trust to the trustee's trustworthiness. That is, one can be generally trustworthy which corresponds to the two-place trust relation. For instance, we might say "Ann is trustworthy". What we are attributing to Ann here is general trustworthiness. At the same time, one can also be narrowly, trustworthy with regard to a particular matter (Jones 1996), which corresponds to the three-place trust relation. For instance, we might say "George is trustworthy when it comes to doing the dishes". What we are attributing to George here is not general trustworthiness but trustworthiness with respect a particular matter, i.e. dishwashing. To keep things simple, we will henceforth refer to the latter property as 'trustworthiness to *phi*' and to the former as 'trustworthiness *simpliciter*'.

Methodologically, focus on one or the other of these two types of trustworthiness has structured the literature in two opposing camps: traditional, demanding views of trustworthiness deal well with trustworthiness *simpliciter* but face several difficulties when it comes to explaining the less high-brow kind of trustworthiness: trustworthiness to *phi*. In response to these problems, more recent work by Katherine Hawley ventures to offer a less demanding account of trustworthiness. Hawley starts her analysis with the three-place, every-day, less highbrow trust relation and the corresponding trustworthiness relation, trustworthiness to *phi*. Unsurprisingly, this view faces difficulties when it comes to satisfactorily accounting for trustworthiness *simpliciter*.

In what follows, we take a quick look at the literature through this lens. It may be worth noting that while we have some new material on Hawley's account, adding to the problems for extant accounts is not our primary concern. Rather our central aim is to

reveal what we will call a demandingness trilemma that accounts of trustworthiness encounter.

2.1 Demanding Accounts: Trustworthiness Simpliciter

Several people working on trustworthiness focus more on trustworthiness *simpliciter*, and, by the same token, on two-place trust. Since the two-place trust relation is intuitively richer, they put forward accounts of trustworthiness that are generally quite demanding.

The classic such account is Annette Baier's (e.g. 1986) *goodwill*-based account (in a similar vein, others combine reliance on goodwill with certain expectations (Jones 1996) including in one case a normative expectation of goodwill (Cogley 2012)). According to this kind of view, the trustworthy person fulfils their commitments *in virtue of* their goodwill towards the trustor. This view, according to Baier, makes good sense of the intuition that there is a difference between trustworthiness and mere reliability, one that corresponds to the difference between trust and mere reliance: trust, but not mere reliance, can be betrayed.

The most widespread worry about these accounts of trustworthiness is that they are too strong: we can trust other people without presuming that they have goodwill (McLeod 2015, REDACTED). Indeed, our everyday trust in strangers falls into this category. In a similar vein, these views give counterintuitive results in the case of two-place trustworthiness: indeed, whether George is trustworthy when it comes to washing the dishes or not seems to not depend on his goodwill, nor on other such noble motives. This seems to suggest that whether or not people are trustworthy out of goodwill is largely inconsequential. Simon Blackburn makes this point forcefully when it comes to trust:

> 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" (Blackburn 1998).

But, of course, there is excellent reason to think that what goes for trust holds, *mutatis mutandis*, for trustworthiness and especially for trustworthiness to *phi*. Otherwise, in the kind of case Blackburn describes here, our trust would have to be misplaced in an important sense. And that doesn't appear to be the case, certainly not in cases in which we trust someone to do something in particular. By the same token, there is reason to think that the goodwill account is too demanding to successfully account for trustworthiness to *phi*.

An alternative to the goodwill account that also focuses on trustworthiness simpliciter is Nancy Potter's view. According to Potter, trustworthiness is a virtue, i.e. a disposition to respond to trust in appropriate ways, given "who one is in relation" to and given other virtues that one possesses or ought to possess (e.g., justice, compassion) (2002: 25). A trustworthy person 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.*" Potter's view purports to account for the intuition that mere reliability is not enough for

trustworthiness by imposing a good character condition on trustworthiness.

In one way, the virtue view is even more demanding than the goodwill view, since good character is a diachronically stable property, while goodwill can be a mere one-off affair. At the same time, the virtue view, is also in a crucial respect also more permissive than the good will view in that it can account for the trustworthiness of strangers insofar as they display the virtue at stake. Their motives on a particular occasion are inconsequential.

Recent criticism of virtue-based views comes from Jones (2012). According to her, trustworthiness does not fit the normative profile of a virtue, in the following way: if trustworthiness were a virtue, then being untrustworthy would be a vice. However, according to Jones, that cannot be right: after all, we are often required to be untrustworthy in one respect or another – i.e. untrustworthy to *phi*; for instance, we can be required to be untrustworthy because of conflicting normative constraints, or untrustworthy when it comes to doing bad things. However, it cannot be that being *vicious* is ever required; therefore, Jones argues, trustworthiness cannot be a virtue. The virtue-based account is also too demanding to work as an account of trustworthiness to phi.[1]

2.2 Hawley's Permissive Account: Trustworthiness to Phi

Katherine Hawley's (2019) new account of trustworthiness departs abruptly from the tradition of taking trustworthiness to be demanding. According to Hawley, trustworthiness is simply a matter of avoiding unfulfilled commitments, which requires both caution in incurring new commitments and diligence in fulfilling existing commitments.

Crucially, on this view, one can be trustworthy regardless of one's motives for fulfilling one's commitments, and regardless of whether one is displaying virtues or not in the process. At the same time, on this view, trustworthiness differs from mere reliability. This is because Hawley accounts for trustworthiness in terms of commitments and people can be reliable *phi*-ers even though the question of commitment to *phi*-ing never arises.

Hawley's is a negative account of trustworthiness, which means that one can be trustworthy whilst avoiding commitments as far as possible. Untrustworthiness can arise from insincerity or bad intentions, but it can also arise from enthusiasm and from becoming over-committed. A trustworthy person, on Hawley's view, must not allow her commitments to outstrip her competence.

One problem for Hawley's account of trustworthiness to *phi* arises from commitments we should have taken on but didn't. For instance, it may well be that Ann should have committed to helping her sick friend George with the groceries last Tuesday but didn't. Now, on the face of it, a person doesn't take on commitments they should have taken on is less trustworthy than someone who does. Or,

[1] Another problem for Potter's view is that defining the trustworthy person as 'a person who can be counted on as a matter of the sort of person he or she is' threatens vicious circularity: after all, it defines the trustworthy as those that can be trusted.

at the very least, a person who never takes on commitments they should have taken on is less trustworthy than a person who always does. But on Hawley's account it is hard to see how this could be. After all, both may be living up to their commitments equally well. It's just that the one person has more commitments than the other.

In response to this problem, Hawley appeals to commitments we take on indirectly, by entering into particular relationships. More specifically, Hawley argues that we may and often do take on meta-commitments — commitments to incur future commitments, by entering into relationships such as friendship, work and other social relationship. As a result, whether or not we take certain new commitments on can matter to how trustworthy we are.

One problem with this move, however, is that it is hard to see how the kind of meta-commitment we incur in virtue of entering into, for instance, a particular friendship could generate a first order commitment so specific as Ann's commitment to help her sick friend George with the groceries last Tuesday. Or, to be more precise, it is hard to see how meta-commitments could generate such specific first-order commitments without generating commitments across the board. And that, in turn, would be hard to square with one of the key motivations for Hawley's negative account, to wit, that one can be trustworthy in virtue of not taking on commitments.

In a similar vein, besides commitments that we should have taken on but didn't, we may undertake bad commitments. To see why this is a problem for Hawley, consider a case in which Ann commits to always lying. It is clear that Ann is not trustworthy when it comes to asserting. At the same time, she may live up to her commitments regarding assertion perfectly. The problem here is that the relevant commitments are bad.

It may also be worth noting that the prospects of explaining what's going on in cases featuring bad commitments in terms of conflicting meta-commitments are dim. One reason for this is that we may simply opt out of the relevant relationship that generates the meta-commitments. Consider Ann, the committed liar. Hawley might say that Ann has a meta-commitment not to lie in virtue of being a member of a community that has a practice of assertion which features a norm against lying. Crucially, however, Ann may opt out of this community. She never believes anything that she is told. Perhaps she even becomes a hermit in order to avoid being exposed to testimony. Here Ann has clearly opted out of any kind of relationship that might generate a meta-commitment against lying. At the same time, she may still be firmly committed to lying should the opportunity arise. If so, she remains an untrustworthy informant.

Third, consider a case in which Ann and George have both committed to meeting a common friend for lunch on ten occasions. Ann lives up to her commitment on all ten occasions. George doesn't make it to two of the ten dates. This is because on one occasion the town he lives in is unexpectedly placed on lockdown, and on the other occasion he gets violently mugged on his way to the lunch place. Since Ann lives up to all of her commitments while George doesn't, Hawley's account predicts that Ann is more trustworthy than George are when it comes to making lunch dates. However, that doesn't seem right. The fact that George doesn't make it on the two

occasions in question doesn't mean that his trustworthiness when it comes to making lunch dates is diminished.

With these problems for Hawley's account of trustworthiness to *phi* in place, let's move on to trustworthiness *simpliciter*. To see why Hawley's account will be too weak to be viable as an account of trustworthiness *simpliciter*, consider first Potter's (2002, 5) case of a sexist employer who treats female employees well only because he believes that he would face legal sanctions if he did not. On Hawley's account, the sexist employer will come out as trustworthy simpliciter; after all, he has taken on a commitment to take his employees fairly and he is making good on this commitment.

Or, consider Ann, our hermit committed liar, once more. Suppose that Ann has not only opted out of social relationship altogether. Moreover, she has become so misanthropic that she now only has bad commitments. Perhaps she has even committed to only having bad commitments. Ann is not a trustworthy person even if she lives up to her commitments.

In this way, there is reason to think that Hawley's negative account of trustworthiness runs into trouble when it comes to accounting for trustworthiness *simpliciter*. Hawley's account makes trustworthiness *simpliciter* too easy to come by.

2.3 A Trilemma for Accounts of Trustworthiness

We have seen that the methodological choice of focusing on different types of trustworthiness has pulled the literature in two opposite directions: accounts that focus their analysis on trustworthiness *simpliciter* end up with rather demanding conditions on trustworthiness. This, in turn, gets them into trouble when it comes to accounting for trustworthiness to *phi*. After all, trustworthiness to *phi* does not plausibly require much in the way of virtue or goodwill on the part of the subject. On the other hand, permissive accounts like Hawley's do better when it comes to trustworthiness to *phi*, but make trustworthiness *simpliciter* too easy to come by. Given the particular methodological foci of these views, this is not surprising. It is, after all, intuitively plausible that trustworthiness *simpliciter* is a demanding affair, in a way in which trustworthiness to *phi* often is not.

By the same token, it looks as though accounts of trustworthiness face the following dilemma:

> An account of trustworthiness is either:
> 1. Demanding: in which case it may successfully account for trustworthiness *simpliciter* but has difficulties accounting for trustworthiness to *phi*
> or
> 2. Permissive: in which case it may successfully account for trustworthiness to *phi* but has difficulties accounting for trustworthiness *simpliciter*

Now, one might wonder whether the dilemma can't be escaped simply by opting for a disjoint approach which takes trustworthiness to *phi* and trustworthiness *simpliciter* to be independent phenomena and offers separate accounts for both.

Unfortunately, there is reason to think that this route is also fraught with problems. To see one, consider first the following conversation:

> **Mary**: People are so untrustworthy these days. Do you know anyone who is actually trustworthy?
> **Ann**: Yes, George. He is really trustworthy.
> **Mary**: How come?
> **Ann**: Well, you can always count on him doing what he's supposed to do. He comes on time to meetings, he finishes all of his work when it's due, he's always there to help his friends and colleagues, and never let me down in the many years since I've known him.

In this exchange, Ann is defending her assessment of George as trustworthy *simpliciter* by citing instances of trustworthiness to *phi* on his behalf. If Ann's reply to Mary is appropriate, which it seems to be, it looks as though the two phenomena are not unrelated. On the contrary, it seems plausible that something like a constitutive relation can be found here, in that trustworthiness *simpliciter* is a function of instances of trustworthiness to *phi*.[2]

Second, consider degrees of trustworthiness *simpliciter*. Compare George with another person James. Suppose James is like George in that he comes to meetings on time, completes his work when it is due, etc. At the same time, James is unlike George in that he isn't always there for friends and colleagues and has let Ann (and other people) down a number of times. It is hard to deny that in virtue of this James is less trustworthy *simpliciter* than George. If trustworthiness to *phi* and trustworthiness *simpliciter* are unrelated phenomena, it is hard to see how this could be. By the same token, there is further evidence that a disjoint approach isn't the way forward either.

What these considerations suggest, then, is the following "demandingness" trilemma for accounts of trustworthiness:

> An account of trustworthiness is either:
> 1. Demanding: in which case it may successfully account for trustworthiness *simpliciter* but has difficulties accounting for trustworthiness to *phi*
> or
> 2. Permissive: in which case it may successfully account for trustworthiness to *phi* but has difficulties accounting for trustworthiness *simpliciter*
> or
> 3. Disjoint: in which case it may successfully account for individual cases of trustworthiness to phi and trustworthiness simpliciter but has difficulties accounting for the relation between the two.

[2] A parallel of this idea, in the literature on trust rather than trustworthiness, is 'three-place-fundamentalism'. According to this view, three-place trust is the fundamental notion and that two-place trust is derivative upon three-place trust (e.g., Baier 1986; Jones 1996; Faulkner 2007; Hawley 2014).

We would like to exclude trying to grasp the third horn of this trilemma from the get-go: a satisfactory account of trustworthiness should be able to account for the two varieties thereof standing in some variety of a close relation to each other.

We also do not favour the route of the first horn: we believe the long history of problematically demanding accounts of trustworthiness constitutes good inductive evidence against going this way. Furthermore, we also think that trustworthiness to *phi* is the more useful concept to clarify, since it tracks the more ubiquitous phenomenon, with most consequences for our everyday life.

In what follows, then, we will attempt to escape the trilemma via the second horn. More specifically, we will develop a new view of trustworthiness that aims to account for both the intuition that trustworthiness *simpliciter* is a rather demanding matter while trustworthiness to *phi* isn't, whilst allowing for a constitutive relation between the two to obtain. In order to do that, we will have a bi-focal approach, whereby we will proceed by discussing the two forms of trustworthiness concomitantly, and by looking into the relation that obtains between them.

3 Trustworthiness: A Bi-Focal Account

We take trustworthiness to *phi* to be more fundamental than trustworthiness *simpliciter*. In what follows, then, we will start off by developing a novel view of the nature of trustworthiness to *phi*. With the account in place, we will turn to the relation that obtains between trustworthiness to *phi* and trustworthiness *simpliciter*. In turn, this will also give us an account of the nature of trustworthiness *simpliciter*.

3.1 Trustworthiness to Phi

The key idea of our view is that trustworthiness to *phi* is in essence a disposition to fulfil one's obligations to *phi*.

Methodologically, we aim to start our analysis with an account of degrees of trustworthiness. In this way, our approach differs from extant approaches in the literature which have ventured to offer accounts of what we will call outright trustworthiness and has featured little to no systematic discussion of degrees of trustworthiness. While this is understandable in that it makes sense to aim to understand what it takes to be trustworthy, the question about degrees of trustworthiness is also an important one. After all, we often want not only to know who is trustworthy and who isn't but also, for instance, who is *most* trustworthy. To understand what it takes to be most trustworthy we need an account of degrees of trustworthiness. Since we offer a systematic account of degrees of trustworthiness, our account promises to fill another important gap in the literature.

How do we make sense of degrees of trustworthiness to *phi*? Our proposal is to start with an account of maximal trustworthiness to *phi*. More specifically, it starts with the following intuitively highly plausible idea: to have the property of trustworthiness to *phi* to its fullest (henceforth also maximal trustworthiness to *phi*) is to as

strongly disposed to fulfil one's obligations to *phi* as possible. Here is the account, more precisely stated:

> **Maximal Trustworthiness to *Phi***
> One is maximally trustworthy with regard to *phi*-ing if and only if one has a maximally strong disposition to fulfil one's obligations to *phi*.

For instance, according to Maximal Trustworthiness to *Phi*, to be maximally trustworthy when it comes to doing the dishes is to have a maximally strong disposition to wash the dishes when under an obligation to wash the dishes.

While we take it to be understood what obligations are and what it takes to fulfil them, we want to say a few words about dispositions. First, dispositions have trigger and manifestation conditions.[3] Consider the disposition of water to boil when heated to 100 degrees centigrade. Here the trigger is heating to 100 degree centigrade and the manifestation is boiling. Similarly, an archer may have the disposition to hit the target upon taking a shot. Here the taking of a shot is the trigger and hitting the target is the manifestation. In Maximal Trustworthiness to *Phi* the trigger is having an obligation to *phi* and the manifestation is fulfilling this obligation to *phi*.

Dispositions are relative to suitable conditions.[4] Water has the disposition to boil when heated to 100 degrees centigrade at but not below sea level. Similarly, an archer may have the disposition to hit the target in normal winds but not in a storm. Dispositions to fulfil obligations are likewise relative to suitable conditions. These conditions may vary depending on the *phi* in question. For instance, George may have the disposition to fulfil his obligations to wash the dishes in his home but not in a place without any water. Suitable conditions for this disposition include the availability of water. At the same time, George may have the disposition to fulfil his obligations to help elderly citizens across the street even in a place without water. Suitable conditions for this disposition do not include the availability of water.

Dispositions may vary in degree of strength. The higher the probability of manifestation given presence of the trigger in suitable conditions, the stronger the disposition.[5] A professional archer has a stronger disposition to hit the target upon taking a shot than we do because the probability of hitting the target conditional on taking the shot whilst being in suitable conditions is higher. The strongest dispositions are dispositions such that the manifestation given the presence of the trigger whilst being in suitable conditions is 1. What Maximal Trustworthiness to *Phi* amounts to is that one is such that the probability that one fulfils an obligation to *phi* given that one has it whilst being in suitable conditions is 1.

---

[3] This is a key marker of dispositionality (McKitrick 2003).

[4] For what we take to be a compelling case that dispositions are relative to suitable conditions, see (Mumford 1998) and (Sosa 2015).

[5] For more on probabilistic approaches to dispositions see (Healey 1991) and (Suarez 2007).

Again, Maximal Trustworthiness to *Phi* is a plausible account of maximal trustworthiness to *phi*: if you have a maximally strong disposition to live up to your obligations to *phi* whilst being in suitable conditions, then you are guaranteed to do. That means that you are maximally trustworthy when it comes to *phi*-ing. To see that the converse holds, note that if someone doesn't have a maximally strong disposition to live up to their obligation to *phi*, their trustworthiness can be improved by strengthening the disposition. [6]

Maximal Trustworthiness to *Phi* states necessary and sufficient conditions for maximal trustworthiness to *phi*. At the same time, we human beings are finite and so we are rarely if ever in the ballpark for maximal trustworthiness to *phi*. Nevertheless, we frequently attribute trustworthiness to *phi* to each other. It would be nice if we could make sense of this practice. How can this be done? And, in particular, how can we make sense of our practice of attributing trustworthiness to *phi* such that at least some of these attributions come out true?

To answer these questions, we'd first like to offer the following account of degrees of trustworthiness to *phi*:

> **Degrees of Trustworthiness to *Phi***
> The degree of trustworthiness to *phi* of S is a function of the distance from maximal trustworthiness to *phi*: the closer one approximates maximal trustworthiness to *phi*, the higher one's degree of trustworthiness to *phi*.

Suppose that while George is generally disposed to live up to his obligation to wash the dishes, he may fail to do so when the Eurovision finals are on or when he is about to finish the book he is

---

[6] One might wonder whether the very idea of maximal trustworthiness isn't problematic because trust essentially involves the incurring of some level of risk of betrayal (McLeod 2015). Maximal trustworthiness would eliminate all risk of betrayal with the result that a person who is maximally trustworthy lies outside the scope of being *trusted*. But that seems implausible because a maximally trustworthy person is surely one whom it is possible to trust.

Three points by way of response. First, we are a little wary about the claim that trust *essentially* involves risk of betrayal. To see why, consider the US's official motto "In god we trust". Presumably, this is compatible with god being such that there is no level of risk of betrayal by god. Champions of the motto are not *ipso facto* committed to the claim that god might betray them.

Second, it is of course true of finite human beings that there is always the possibility of risk of betrayal. After all, finite human beings are fallible creatures living in a complex world with many and often enough conflicting normative constraints. So, the claim may still be true of finite human beings.

Third, even if there the risk of betrayal can be eliminated on the side of the trustee, it may be that there is a risk of betrayal cannot be eliminated on the side of the trustor. For instance, if the trustee is god, it may be impossible that you will be betrayed. At the same time, if you have no idea that the trustee is god, you may be unable to eliminate the risk of betrayal. But that should be enough to support the possibility of trusting someone who is maximally trustworthy, even by the lights of advocates of the above worry.

reading. Ann is also generally disposed to live up to her obligation to wash the dishes. She may fail to do so when the Eurovision finals are on, but she will not let an almost finished book get in the way. Degrees of Trustworthiness to *Phi* predicts that Ann is more trustworthy when it comes to doing the dishes than George is. Since that's the right result intuitively, the news is good for our account on this front.

Next, we would like to combine this account of degrees of trustworthiness to *phi* with a contextualist semantics for outright attributions of trustworthiness to *phi*. According to our account of outright attributions of trustworthiness to *phi*, context determines a threshold on degrees of trustworthiness to *phi* such that one is trustworthy to *phi* just in case one surpasses the threshold in question. Or, to be more precise,

> **Attributions of Outright Trustworthiness to *Phi***
> "*S* is trustworthy to *phi*" is true in context *c* if and only if *S* approximates maximal trustworthiness to *phi* closely enough to surpass a threshold on degrees of trustworthiness determined by *c*.

On this view, then, when, at a particular context, we say that Ann is trustworthy when it comes to washing the dishes but George isn't, what is happening is that Ann approximates a maximally strong disposition to do so, conditional on having the corresponding obligation, to a contextually sufficiently high degree, whereas George doesn't. To return to our example, Ann may fail to live up to her obligation to wash the dishes once a year, if her turn in the rota falls on the night of the Eurovision finals. Her disposition to live up to her obligation to wash the dishes is very strong, and very plausibly strong enough to take her close enough to the maximum to surpass the threshold and make the attribution of trustworthiness come out true. At the same time, suppose George finishes a book every three days and it's his turn to wash the dishes every other day. In that case, his disposition to live up to the obligation in question is considerably less strong and may very well not be strong enough to take him close enough to the maximum to make the threshold. In that case, he cannot truly be attributed trustworthiness when it comes to doing the dishes.

Finally, note that the semantics is contextualist in that just how high the threshold is will be determined by and may vary with context. To see that this is plausible, compare a case in which Ann and George are professional dishwashers at a local restaurant with a case in which they are Mary's teenage children. It is intuitively plausible that the threshold for what it takes to count as trustworthy when it comes to washing the dishes is higher in the first case than in the second. Since our account can accommodate this intuition, this is again good news for it.

3.2 Trustworthiness Simpliciter

We have noted above on several occasions that, while the relation between trustworthiness *simpliciter* a trustworthiness to *phi* is not straightforward at all, there is one thing that seems plausible from

the get-go, to wit, that trustworthiness *simpliciter* will be, in one way or another, constituted by a set (to be specified) of instances of trustworthiness to *phi*. This plausible idea is at the very heart of our account of trustworthiness *simpliciter*. In what follows, we will it flesh out in more detail, thereby developing our account of trustworthiness *simpliciter*.

Recall that we said that trustworthiness comes in degrees and that we started our analysis of trustworthiness to *phi* with an account of maximal trustworthiness to *phi*. It will not be surprising, that we follow the same approach for our account of trustworthiness *simpliciter*. Once again, we take our lead from an intuitively highly plausible thought. Here it is: to have the property of trustworthiness *simpliciter* to its fullest (henceforth also maximal trustworthiness *simpliciter*) is to be maximally trustworthy to *phi* on all counts. Here is the account, more precisely stated:

> **Maximal Trustworthiness *Simpliciter***
> One is maximally trustworthy simpliciter if and only if one is maximally trustworthy to *phi* for all *phi*.

Maximal Trustworthiness *Simpliciter* states necessary and sufficient conditions for maximal trustworthiness simpliciter. However, again, we, human beings, are finite and so we are rarely if ever in the ballpark for maximal trustworthiness. At the same time, we frequently attribute trustworthiness *simpliciter* to each other. It would be nice if we could make sense of this practice. Moreover, it would be nice if we could make sense of this practice such that some of these attributions can come out true. How can this be done?

Again unsurprisingly, the answer we would like to propose combines an account of degrees of trustworthiness with a contextualist semantics for outright attributions of trustworthiness simpliciter. Given that Maximal Trustworthiness *Simpliciter* gives us the maximum degree of trustworthiness, it will come as no surprise that our account of degrees of trustworthiness measures degrees of trustworthiness in terms of approximations to the maximum degree:

> **Degrees of Trustworthiness Simpliciter**
> The degree of trustworthiness *simpliciter* of S is a function of the distance from maximal trustworthiness to *phi* for all *phi*: the closer one approximates maximal trustworthiness to *phi* for all *phi*, the higher one's degree of trustworthiness *simpliciter*.

For instance, suppose that Ann is the most trustworthy human being imaginable: she is trustworthy when it comes to doing her work well, when it comes to helping her friends, when it comes to doing the dishes and so on. George is just like Ann with one exception: he is not equally trustworthy when it comes to doing the dishes (perhaps for the reasons mentioned above). According to Degrees of Trustworthiness *Simpliciter*, Ann is more trustworthy *simpliciter* than George. Since that is the intuitively correct result, the news for our account is once again good.

Regarding the contextualist semantics for outright attributions of trustworthiness, the key idea of our proposal is that an attribution of outright trustworthiness to one is true just in case one approximates maximal trustworthiness to *phi* for all *phi* closely enough. How close is close enough? As is commonly the case with gradable expressions, the answer is that this depends on contextual factors. In other words, context determines a threshold of distance to maximal trustworthiness that one must surpass for outright trustworthiness to be truly attributable to one.

> **Attributions of Outright Trustworthiness *Simpliciter***
> "*S* is trustworthy" is true in context *c* if and only if *S* approximates maximal trustworthiness to *phi* for all *phi* closely enough to surpass a threshold on degrees of trustworthiness determined by *c*.

According to this view, then, when Ann says "George is trustworthy", what she is saying is that he is trustworthy enough for the conversational context.

3.3 Threshold Setting

Now, note that degrees of trustworthiness *simpliciter* can be measured along at least two dimensions, i.e. breadth and depth: we can measure on how many *phi*-s one is trustworthy on, and how well one approximates maximal trustworthiness to *phi* for each *phi* in question. In turn, both of these dimensions will influence how the threshold is set at a given context. How so?

We will start with the latter dimension – i.e. depth – since our view affords an easy answer to the threshold question in this regard: after all, we have proposed that "*S* is trustworthy to *phi*" is true in context *c* if and only if *S* approximates maximal trustworthiness to *phi* closely enough to surpass a threshold on degrees of trustworthiness determined by *c*. This, together with our account of trustworthiness in terms of dispositions to fulfil one's obligations, gives us the result that the depth dimension of the contextual threshold for trustworthiness simpliciter, which is given by the contextually appropriate degree of trustworthiness to *phi* for a particular *phi,* concerns the contextually appropriate strength of one's disposition to meet one's obligations to *phi*.

How about the threshold for breadth? This is more complicated. Our proposal is that the breadth threshold is to be understood in terms of a contextually determined set (or sets) of *phi*;[7] that is, the set (or sets) of actions that are salient at the conversational context where the attribution is made.

To see this, it will be useful to take a short detour and distinguish between two varieties of ascriptions of trustworthiness *simpliciter*, viz. predicative and attributive:

[7] We want to allow that one can be trustworthy *simpliciter* in different ways, i.e. by approximating maximal trustworthiness *simpliciter* via different routes, as it were. To do achieve this, we may countenance sets of sets of phi-ings that are made salient such that one is close enough to maximal trustworthiness *simpliciter* just in case one is sufficiently trustworthy to *phi* for all *phi* in some such set of sets.

> Predicative ascriptions: George is trustworthy. Ann is trustworthy. Attributive ascriptions: George is a trustworthy babysitter. Ann in a trustworthy physician. [8]

Let's first look at attributive ascriptions of trustworthiness *simpliciter* as in George is a trustworthy babysitter and Ann is a trustworthy physician. How is the threshold for breadth determined in cases of attributive ascriptions of trustworthiness *simpliciter*? As a first step, the relevant *phi*-s that are picked up by the conversational context are the *phi*-s pertaining to the domain of attribution, i.e. babysitting in the case of George (watching the kids, feeding them, etc.), and being a physician in the case of Ann (diagnosing health conditions, prescribing medication, etc.).

That said, practical interests of the attributors also matter. Consider two cases in which the question whether Ann is a trustworthy physician is under consideration. In one case, the question is asked by the owners of the hospital at which Ann works. In the other case, the question is asked by prospective patients. Note that it may well be that "Ann is a trustworthy physician" can be naturally and intuitively correctly asserted in one context but not in the other. To see this, suppose Ann will deviate from lawful procedure when she deems it adequate to restoring her patient's health. In that case, it might well be that in the context at issue in the case of the patients "Ann is a trustworthy physician" is natural and intuitively correct. At the same time, in the context at issue in the case of the hospital owners "Ann isn't a trustworthy doctor" is natural and intuitively correct. If we allow for practical interest to play a part in determining the set of *phi*-s relevant to making the threshold for breadth, we can easily accommodate these intuitions. In the context at issue in the case of the patients, practical interests determine a set of *phi*-s that doesn't include following lawful procedure such that "Ann is a trustworthy doctor" comes out true. At the same time, in the context at issue in the case of the hospital owners, practical interests determine a set of *phi*-s that does include following lawful procedure such that "Ann isn't a trustworthy doctor" comes out true.

The way the threshold for attributive ascriptions of trustworthiness simpliciter is set at a context, then, is as follows: first, context delivers the *phi*-s that are relevant for the breadth dimension of the threshold against which the trustworthiness ascription is to be evaluated as true or false in accordance with the phi-s pertaining to the domain of attribution and practical interests. Second, after the set (or sets) of *phi*-s is established, the threshold for depth across the *phi*-s in question gets set: that is, in this second step, context determines how strong the disposition to fulfil one's obligations to *phi* for the relevant *phi*-s needs to be.

What about predicative ascriptions of trustworthiness simpliciter, i.e. when by ascribing trustworthiness *simpliciter* we mean to ascribe a more general property? For instance, suppose we say "George is trustworthy" and we mean to say that George is generally to be trusted. What is happening in these cases?

---

[8] For more on predicative vs. attributive ascriptions see (Geach 1956).

There are two options. First, these cases are also implicitly attributive. Alternatively, second, they are not. Rather, they are ascriptions of trustworthiness *simpliciter* that are what we will call "purely predicative". Note that we can easily make sense of the second option. What it takes to count as trustworthiness *simpliciter* is simply to be close enough trustworthiness to *phi*, for all *phi*. While there is no restriction on breadth via domain of attribution, there may still be a restriction via practical interest. As a result, our account can accommodate the intuition that utterances of "George is trustworthy" may have different truth conditions depending on whether it they are made in discussions among parents or teenagers, for instance.

If there are purely predicative ascriptions of trustworthiness simpliciter, the threshold for breadth is set at a context in the following way: first, context delivers the *phi*-s that are relevant for the breadth dimension of the threshold against which the trustworthiness ascription is to be evaluated as true or false in accordance with the practical interests of the attributors. Second, after the set (or sets) of *phi*-s is established, the threshold for depth across the *phi*-s in question gets set: that is, in this second step, context determines how strong the disposition to fulfil one's obligations to *phi* for the relevant *phi*-s needs to be.

While our account can accommodate the possibility of purely predicative ascriptions of trustworthiness simpliciter, there is also reason to think that there are no purely predicative ascriptions. Instead, all predicative ascriptions of trustworthiness simpliciter are implicitly attributive. To see this, consider Ann and George, two persons, and Starbucks and Costa, two corporations. Note that the following sound perfectly fine:

1. Ann and George are equally trustworthy.
2. Starbucks and Costa are equally trustworthy.

In contrast, the following sounds odd:

3. ??Ann and Starbucks are equally trustworthy.[9]

If there are purely predicative ascriptions of trustworthiness, it is hard to see why (1) and (2) sound fine, while (3) sounds odd. On the other hand, if all predicative ascriptions of trustworthiness simpliciter are all implicitly attributive, we may be able to do better.

Here is our proposal: gradable adjectives such as "tall", "heavy" and "trustworthy" have a dimension that corresponds to a gradable property – height, weight and trustworthiness – and allows ordering of sets of objects. "Equally" is a presupposition trigger for a common dimension. For instance, when we say "Ann and George are

[9] Compare also: "London and Lima are equally large" and "The ratio of people over 65 to people under 65 in the UK and the ratio of people earning more than GBP50000 to people earning less than GBP50000 in the UK" are equally large" both sound perfectly fine. However, "Lima and the ratio of people over 65 to people under 65 in the UK are equally large" sounds odd. The phenomenon we are pointing to here is thus not isolated.

equally tall" we are presupposing that there is a common dimension, here the one corresponding to height.[10]

Now, if all predicative ascriptions of trustworthiness *simpliciter* are all implicitly attributive, this presupposition may turn out to be false. After all, it may be that there is no common dimension that allows for ordering of sets of objects. Rather, it may be that all we have is two different dimensions one that corresponds to trustworthiness in Xs and the other to trustworthiness in Ys. Now suppose this turns out to be the case for (3). In that case, it is unsurprising that (3) should sound odd. After all, (3) carries a false presupposition.

This leaves the question as to what it is about trustworthiness such that there is no common dimension for "Ann and Starbucks are equally trustworthy." Here is our proposal: What we say when we say that George is trustworthy *simpliciter* in the most general sense – i.e. in cases that are the best candidates for purely predicative ascriptions of trustworthiness *simpliciter* – is just that he is a trustworthy member of the most general kind of which he is a member and that has obligations associated with it.[11] In the case of Ann and George, the kind is the kind 'person.' In the case of Starbucks and Costa Coffee, it is the kind 'corporation.' While there are more general kinds of which both Ann and Starbucks are members, e.g. the kind entity, this kind is not a kind that has obligations associated with it. There are no obligations one incurs simply in virtue of being an entity. Since there is no kind that has obligations associated with it such that Ann and Starbucks are both members of this kind, if predicative ascriptions of trustworthiness *simpliciter* are always implicitly attributive in the way envisaged, it follows that there is no common dimension of trustworthiness for Ann and Starbucks. As a result, if we agree that predicative ascriptions of trustworthiness simpliciter are always implicitly attributive, we can explain why (3) sounds odd. At the same time, it is easy to see why (1) and (2) don't sound odd. Here there the presupposition in question isn't false. After all, Ann and George are both persons and Starbucks and Costa are both corporations.[12]

---

[10] Note that when we say "Ann and George aren't equally tall", "Are Ann and George equally tall?" we are also presupposing that there is a common dimension of height. Moreover, "Ann and George are equally tall, but they don't have a height" sounds odd. This indicates that "equally" passes the standard diagnostic tests for presuppositions, i.e. projectability and non-cancellability when the trigger is not embedded (Beaver and Geurts 2014).

[11] More specifically, there is reason to believe that being a member of a kind with associated obligations is also a presupposition triggered by "trustworthiness". To see this, consider "Baby Joe is trustworthy" and "Baby Joe isn't trustworthy". Both sound odd. The proposed presupposition can explain why: Baby Joe is not a member of a kind with associated obligations, at least not yet. The presupposition is false.

[12] But can't we felicitously say things like "Facebook is more trustworthy than Donald Trump" or "89% of Republications think Donald Trump is more trustworthy than CNN"? At the same time, if there is no common dimension for corporations and persons, it is hard to see how these claims could be felicitous. In response, we want to agree that there is a common dimension here: being a trustworthy source of information. Strictly speaking, then, it's not the case that there is no common dimension for

In contrast, if we allow that there are purely predicative ascriptions of trustworthiness simpliciter, this explanation will not be available to us. After all, in that case, the presupposition of a common dimension for Ann and George is true. An explanation of why (3) sounds odd that appeals to the falsity of this presupposition is not on the cards. This is not to say that there might not be another explanation. However, it does place the onus squarely on those who think that there are purely predicative ascriptions of trustworthiness simpliciter.

4 Comparison with the Competition

With our account of trustworthiness on the table, we will now return to the problems that the competition faces and argue that our account does better. More specifically, we will argue that not only can our account avoid the problems that its rivals encounter, but it can also accommodate their most important insights.

4.1 The Trilemma

First and foremost, our account promises to escape the trilemma we sketched in Section 2.

Since our approach is bi-focal it can offer accounts of not only of trustworthiness *simpliciter* and trustworthiness to *phi*, but also of the relation between the two. In particular, the account explains trustworthiness *simpliciter* in terms of trustworthiness to *phi*, i.e., roughly, in terms of trustworthiness to *phi* across the board. In this way, it offers an attractive account of the relation between the two kinds of trustworthiness and avoids the drawbacks of a disjoint account.

At the same time, as we will argue momentarily, it can improve on the competition in the literature in that it can avoid construing trustworthiness *simpliciter* as demanding enough to be adequate only at the expense of offering too demanding an account of trustworthiness to *phi*. In addition, it can also avoid construing trustworthiness to *phi* as permissive enough to be adequate only at the expense of making trustworthiness *simpliciter* too easy to come by. Let's look at the details.

4.2 The Goodwill Account

Recall that the goodwill account struggles with trustworthiness to *phi*. Someone may be trustworthy when it comes to doing the dishes even if they have little to no goodwill. While this is indeed problematic if goodwill is required for trustworthiness, it is easy to see that our account can avoid this problem. What matters to

---

corporations and persons. At the same time, context doesn't always make a common dimension salient. In the Trump cases it does, in (3) it doesn't. The fact that we find trustworthiness comparisons between persons and corporations odd when context doesn't provide a common dimension still provides evidence for the claim that predicative ascriptions of trustworthiness are implicitly attributive. After all, it's plausible that the presupposition triggered by "equally" is not only that there is some common dimension, even if we have no clue as to what it may be, but also that it is contextually salient. And that presupposition is still false for (3) but not for (1) and (2).

trustworthiness to *phi* is a disposition to fulfil one's obligations to *phi*. And it is of course entirely possible to possess this particular disposition without having goodwill. As a result, our account avoids this problem.

Does this mean that goodwill has nothing to do with trustworthiness at all on our account? No. To see why not, consider trustworthiness *simpliciter*. Note that we often have obligations not only to *phi* but to *phi* for the right reason. A person with goodwill will have a stronger disposition to *phi* for the right reason than a person without. On our account, then, the person with goodwill has a higher degree of trustworthiness *simpliciter* than a person without goodwill.

Relatedly, it is worth noting that for human beings, it is hard to see what might ground a disposition to fulfil our obligations to *phi*, for a wide range of *phi*, except goodwill.[13] By the same token, it is hard to see how normal human beings could attain certain degrees of trustworthiness *simpliciter* unless they also have goodwill. And, of course, this means that there will be contexts in which it is hard to see how ascriptions of trustworthiness *simpliciter* might come out true of people unless they also possess goodwill.

Note also that our account can preserve the distinction between trustworthiness and reliability. There are at least a couple of ways in which the two can come apart. First, consider a person who simply doesn't have the obligation to *phi*. We may still rely on this person to *phi*, when they have the disposition to *phi* whether or not they have the obligation to do so. For instance, if Ann has a disposition to buy her morning coffee at the local coffee shop, it may make sense for us to rely on her doing so. At the same time, since she doesn't have an obligation to buy your coffee at the local coffee shop, trustworthiness doesn't enter the picture in the first place. It may also be worth noting that, as a result, when it turns out that one morning Ann didn't buy her coffee at the local coffee shop, while we may be disappointed, we are not entitled to feel betrayed. This makes perfect sense, given that Ann didn't have an obligation to do so in the first place.

Another kind of case in which reliance and trustworthiness come apart are cases in which there is a mimicker for the disposition to fulfil one's obligation. Mimickers bring about the manifestation of a disposition when the trigger condition obtains in things that don't have the disposition. The classical example is the case of the hater of Styrofoam. Styrofoam has the disposition to produce a distinctive kind of sound when struck. It does not have the disposition to break when struck. Now suppose that there is a hater of Styrofoam who will show up whenever he hears the distinctive kind of sound of striking a piece of Styrofoam and tears the Styrofoam apart. In this case, Styrofoam still doesn't have the disposition to break when struck. Rather the hater of Styrofoam is a mimicker for this disposition (Lewis 1997).

There may be mimickers of the disposition to fulfil one's obligations. Suppose that whenever George has an obligation to wash

[13] This is not to say that other ways are inconceivable. However, these other ways are fanciful and do not plausibly obtain widely in normal human beings.

the dishes, an army of fairy helpers sees to it that the dishes get done when George is too lazy to do it. In this case, we can rely on the dishes getting done when it's George's turn to do them. At the same time, George may well be highly untrustworthy when it comes to doing the dishes. Again, we have a case in which reliability and trustworthiness come apart. It may be worth noting that, in contrast with the case of Ann who doesn't have the relevant obligation, in this case, when we discover what's going on, we are still entitled to feel betrayed even if we can continue to rely on George. Once again, this makes perfect sense since George had the obligation to wash the dishes but didn't fulfil it.

4.2 The Virtue Account

Let's move on to the virtue account then. Here the central problem was that trustworthiness cannot be a virtue. This is because we can never be required to be vicious. At the same time, there are cases in which we can be required to be untrustworthy for instance in cases of normative conflict and in cases of doing bad things.

It is easy to see that, on our view, trustworthiness isn't a virtue. In particular, trustworthiness to *phi* isn't a virtue. To see this, recall the case of partners in crime. One partner in crime may be trustworthy when it comes to not giving the other one up to the police. However, that doesn't make them virtuous. In consequence, trustworthiness to *phi* can't be a virtue.

While trustworthiness to *phi* isn't a virtue, the question remains whether trustworthiness *simpliciter* is a virtue. The answer to this question will turn on the relationship between virtues, norms and dispositions. We will not venture to answer it here. We will, however, say this much. If it turns out that trustworthiness *simpliciter* is a virtue, there is no problem from cases in which we are required to be untrustworthy *simpliciter*. Here is why. If trustworthiness simpliciter is a virtue, we should be trustworthy *simpliciter*. In other words, we have an obligation to be trustworthy *simpliciter*. This already makes it somewhat hard to see how we could ever be required (i.e. be obligated) to be untrustworthy *simpliciter*. Of course, if we can't, there is no problem from cases in which we are required to be untrustworthy *simpliciter*.

Now suppose this is too quick and there are scenarios in which we are required to be untrustworthy *simpliciter*. What might a case like that look like? Here is one example. Suppose someone with immense powers tells Ann that they will end humanity unless Ann becomes untrustworthy *simpliciter*. In this case, Ann may be required to be untrustworthy *simpliciter*. More generally, cases in which we are required to be untrustworthy *simpliciter* are cases in which the obligation to be trustworthy *simpliciter* is overridden by a different obligation. But even so, in this case, either being untrustworthy *simpliciter* is compatible with not being vicious or else we are required not only to become untrustworthy *simpliciter* but also to become vicious. Either way, the problem from cases in which we are required to be untrustworthy *simpliciter* can be avoided.

4.3 Hawley's Commitment Based Account

Recall Hawley's account encounters problems on four fronts. First, it ran into trouble with commitments we should have taken on. Second, cases of bad commitment were problematic. Degrees of commitment were a third source of difficulty. And, fourth, Hawley's account also struggled with trustworthiness *simpliciter*.

It is easy enough to see that our account can avoid the first problem. After all, in cases in which we should take on a commitment, we have an obligation to take on the commitment. To the extent that we aren't disposed to fulfil this obligation, our trustworthiness is diminished. To be more precise, our trustworthiness to take on commitments we should have taken on is diminished. What's more, while it is in principle possible to have a weak disposition to take on commitments to *phi* when one should and yet to have a strong disposition to *phi* when one should, in practice, in the vast majority of cases the weakness of the former disposition is grounded in the weakness of the latter disposition. In these cases, of course, what is diminished is not only our trustworthiness when it comes to taking on the commitments to *phi* that we should take on but also our trustworthiness when it comes to *phi*-ing itself. In this way, our account avoids the problem.

The second problem concerns bad commitments. Our toy case here featured Ann, the hermit committed liar. While Ann lives up to her commitment perfectly, she is untrustworthy when it comes to asserting. Again, our account improves on Hawley. The fact that Ann commits to only ever asserting lies does not mean that the norm that prohibits asserting lies doesn't apply to her. At the same time, not all commitments generate obligations. In fact, Ann's commitment to only ever assert lies doesn't generate such a corresponding obligation. What happens in the case in which Ann commits to only ever asserting lies, is that she loses any disposition to fulfil their obligation not assert lies she may have had before. By the same token, she comes out as not trustworthy when it comes to not asserting lies.

What about the problems for degrees of trustworthiness Hawley encounters? Recall the case in which Ann lives up to her commitment and makes all ten lunch dates, but George only makes eight because his town goes into lockdown on one occasion, and he gets violently mugged on the other. Hawley's account predicts that Ann is more trustworthy than George. However, clearly, the fact that George doesn't live up to his commitments through no fault of his own shouldn't diminish his trustworthiness when it comes to making lunch dates.

To see how our account deals with this problem, recall that we take dispositions to be relative to suitable conditions. The fact that suitable conditions don't obtain doesn't diminish the strength of one's disposition relative to those conditions here. Rather, what's going on here is that the disposition is masked (e.g. Johnston 1992, Bird 1998). And this is what explains why George's trustworthiness when it comes to making lunch dates isn't diminished here. What's happening is that suitable conditions don't obtain. George's disposition to make lunch dates is masked and its strength is not affected. As a result, the fact that George doesn't live up to his

commitment to make the lunch date doesn't negatively affect his degree of trustworthiness.

Let's look at the last problems that Hawley encounters which concern trustworthiness *simpliciter*. The sexist employer who lives up to their commitment to treat their female employees fairly will of course come out as trustworthy when it comes to treating female employees fairly, at least when they have the corresponding disposition. However, whether they are trustworthy *simpliciter* is a different question. In particular, there are obligations not to be sexist and to treat employees fairly for the right reason. Our sexist employer is not disposed to fulfil these obligations which diminishes their degree of trustworthiness *simpliciter*.

And regarding Ann, the hermit committed liar, who has become so misanthropic as to only have bad commitments and indeed is committed to only ever taking on bad commitments, we can again point out that her bad commitments don't absolve her of the obligations she has and that not all commitments generate obligations. As a result, when Ann ends up with only bad commitments and a commitment to only ever take on bad commitments, she doesn't have a disposition to fulfil her obligations to *phi*, for a wide range of (perhaps all) *phi*. On our view, then, he scores low on trustworthiness *simpliciter*.

Before closing, we want to briefly consider the relation between commitments and trustworthiness. Just as in the case of goodwill, we may wonder whether this relation is severed entirely. Again, the answer is no. The reason for this is that taking on commitments typically means taking on obligations. For instance, when we commit to going for lunch with a friend, we have now an obligation to do so. As a result, to live up to one's commitments typically means fulfilling the corresponding obligation. We don't mean to deny that there is an important relationship between commitments and trustworthiness. More specifically, we want to allow that how well one lives up to one's commitments may well be a decent measure of trustworthiness.[14]

5 Conclusion

We have undertaken a methodological turn in this paper, in that we approached the issue of the nature of trustworthiness bi-focally: by focusing on the *relation* between trustworthiness *simpliciter* and trustworthiness to *phi*, rather than on one or the other of these two phenomena. In turn, the account we put forth and defended – according to which trustworthiness is a disposition to fulfil one's obligations – successfully escapes the demandingness trilemma: it accounts, at the same time, for the intuitively highbrow nature of trustworthiness simpliciter, for the fact that two place trustworthiness is easy to come by, and for the existence of a constitutive relation between them.

[14] That said, it is important to keep in mind that the prospects for an account of trustworthiness in terms of commitments are dim, for the reasons mentioned above.

References

Baier, A. 1986. Trust and Antitrust. *Ethics* 96: 231–260.

Beaver, D. and Geurts, B. 2014. Presupposition. In Zalta, E. ed. *The Stanford Encyclopedia of Philosophy*. URL = <https://plato.stanford.edu/archives/win2014/entries/presupposition/>.

Bird, A. 1998. Dispositions and Antidotes. *The Philosophical Quarterly* 48: 227–234.

Blackburn, S. 1998. *Ruling Passion: A Theory of Practical Reasoning*. Oxford: Clarendon Press.

Cogley, Z. 2012. Trust and the Trickster Problem. *Analytic Philosophy*, 53: 30–47.

Faulkner, P. 2007. A Genealogy of Trust, *Episteme*, 4(3): 305–321.

Geach, P. 1956. Good and Evil. *Analysis* 17: 33–42.

Hawley, K. 2019. *How to Be Trustworthy*. Oxford: Oxford University Press.

Healey, R. 1991. *The Philosophy of Quantum Mechanics: An Interactive Interpretation*. Cambridge: Cambridge University Press.

Johnston, M. 1992. How to Speak of the Colors. *Philosophical Studies* 68: 221–263.

Jones, K. 1996. Trust as an Affective Attitude. *Ethics* 107: 4–25.

Lewis, D. 1997. Finkish Dispositions. *The Philosophical Quarterly* 47: 143–158.

McKitrick, J. 2003. A Case for Extrinsic Dispositions. *Australasian Journal of Philosophy* 81: 155–174.

Mumford, S. 1998. *Dispositions*. Oxford: Oxford University Press.

McLeod, C. 2015. Trust. *The Stanford Encyclopedia of Philosophy*. In Zalta, E. ed., URL = <https://plato.stanford.edu/archives/fall2015/entries/trust/>.

Potter, N. 2002. *How Can I be Trusted? A Virtue Theory of Trustworthiness*. Lanham, Maryland: Rowman & Littlefield.

Sosa, E. 2015. *Judgment and Agency*. Oxford: Oxford University Press.

Suarez, M. 2007. Quantum Propensities. *Studies in History and Philosophy of Modern Physics* 38: 418–38.