Friendship and the Structure Of Trust

Mark Alfano

1. Introduction

Friendship might seem like a bizarre virtue – or not a virtue at all. In Aristotle's early discussion of the virtues in the *Nicomachean Ethics*, we see courage, temperance, generosity, magnificence, magnanimity, pride, wit, and justice. These would all seem to be, in the first instance and primarily, monadic properties of individual agents. To be courageous is to be disposed to think, feel, desire, deliberate, act, and react in characteristic ways. Even if no one else is courageous, it would still be possible – though extremely difficult – for you to be courageous. Of course, if there are no threats to be opposed, you may never have a chance to manifest or express your courage. Furthermore, it would surely be easier to develop courage in the company of others who either are or strive to be courageous. And it may also be easier to develop or sustain courage when others think of you as courageous and signal those thoughts to you. But, one might think, even if none of these enabling facts obtains, it would still be possible, conceptually speaking, to be courageous.

Or consider generosity. To be generous is to be disposed to think, feel, desire, deliberate, act, and react in characteristic ways. Even if no one else is generous, it would still be possible – though extremely difficult – for you to be generous. Of course, if there were no other people who needed or wanted or would appreciate what you have, or if you were so down on your luck that you had no resources to offer, you may never have a chance to manifest or express your generosity. Furthermore, it would surely be easier to develop generosity in the company of others who are or strive to be generous (and, for that matter, grateful). And it may also be easier to develop or sustain generosity when others think of you as generous and signal those thoughts to you. But, one might think, even if none of these enabling facts obtains, it would still be possible, conceptually speaking, to be generous.

Friendship appears to be different. It seems to be, in the first instance and primarily, a dyadic relation between two people. To be a friend is to be disposed to think, feel, desire, deliberate, act, and react in characteristic ways towards a particular person, who is likewise disposed to think, feel, desire, deliberate, act, and react in those same characteristic ways towards you. If no one else is a friend, then it is conceptually impossible – not just difficult – for you to be a friend as well (*NE* VIII:2, 1155a; see Cooper 1977b, p. 624 and Nehamas 2010, p. 216). It is not just *easier* to develop friendship in the company of others who are doing so as well; it is in fact impossible to become a friend without there being someone else who also becomes a friend, namely *your* friend.

In this paper, I describe some of what I take to be the more interesting features of friendship, then explore the extent to which other virtues can be reconstructed as sharing those features. I use trustworthiness as my example throughout, but I think that other virtues such as generosity & gratitude, pride & respect, and the producer's & consumer's sense of humor can also be analyzed with this model. The aim of the paper is not to demonstrate that all moral virtues are exactly like friendship in all important respects, but rather to articulate a fruitful model in which to explore the virtues. Section 2 explores the relational nature of friendship, drawing on Aristotle's discussion of friendship in the *Nicomachean Ethics*. Section 3 catalogues four motivations for taking seriously the friendship model of virtue. Section 4 applies the friendship model in depth to the virtue of trustworthiness.

2. The Relational Nature of Friendship

Being a friend isn't just a matter of your first-order cognitive, affective, evaluative and behavioral dispositions; to be a friend means, among other things, to have particular *de re* attitudes towards another person (your friend), and for that person to have congruent *de re* attitudes towards you (*NE* VIII:2, 1156a). That is, for you to be my friend, you need to think of me as your friend, to wish me well for my own sake, to wish me well in virtue of my good character (or, in other types of friendship, in virtue of my contributing to your utility or pleasure), and so on. Likewise, I need to think of you as my friend, to wish you well for your own sake, to wish you well in virtue of your good character (or in virtue of your contributing to my utility or pleasure), and so on.

But that is not enough. Not only must both you and I have these attitudes, but the existence of these attitudes must be mutual knowledge between us (*NE* IX:5, 1166b). If I wish you well for your own sake and in virtue of your good character, and you wish me well for my own sake and in virtue of my good character, but neither of us knows how the other feels, we are not friends. Instead, we merely harbor mutual but unrecognized good will towards one another. To be your friend, I need to know that you wish me well for my own sake and in virtue of my good character, and you need to know that I wish you well for your own sake and in virtue of your good character (*NE* VIII:3, 1156b).

In fact, even that is not enough. We could satisfy this description and yet still not be friends. If we have these attitudes towards each other, and each finds out through reliable testimony that the other does as well, it would still seem strange to say that we are friends. We might never have met each other, yet satisfy these conditions. Friends have a more intimate connection than this. I also need to know that you know that I wish you well for your own sake and in virtue of your good character, and you need to know that I know that you wish me well for my own sake and in virtue of my good character.

It's arguable that even this is not enough, and that what needs to hold is that we share common knowledge of our attitudes: I know that you wish me well for my own sake and in virtue of my good character; and you know that I know that you wish me well for my own sake and in virtue of my good character; and you know that I know that you know that I know that you wish me well for my own sake and in virtue of my good character, and so on. Or, somewhat less strongly, perhaps what's required is that there be what Lewis (2002, p. 56) calls a *basis for common knowledge* between us, even if only two levels of mutual knowledge are actually present. I will not press this point here, for even if all that's required is two orders of mutual knowledge (I know that you know, and you know that I know), my point still holds that friendship is an interesting virtue because it requires reciprocated *de re* attitudes and some kind of mutual recognition of the existence of this reciprocation.

Finally, friends typically harbor other, more complicated, attitudes towards one another, and react with higher-order attitudes to the presence or absence of lower-order attitudes. Roberts (2013) argues that *de re* emotional interactions are constitutive of friendship (p.

141); he explores the ways in which emotions and emotional feedback loops strengthen and desiccate such relationships as friendship, enmity, civility, and incivility. For example, consider a sister who generously and in a spirit of friendship gives her brother her own ticket to a concert that he would like to attend. He feels the emotion of gratitude for this gift, which he expresses with a token of thanks. Satisfied that her generosity has hit its mark, she is "gratified by his gratitude. […] And he may in turn be gratified that she is gratified by his gratitude" (p. 137). Despite the fact that this is a tiny schematic example, it plausibly contains a fourth-order emotion (he is gratified that she is gratified that he is gratified that she was generous). Such episodes are, in Roberts's view, constitutive of friendship and other normative personal relationships (pp. 140-1). Constructive feedback loops strengthen positive personal relationships but aggravate negative relationships such as enmity (leading enemies to hate, despise, or contemn each other all the more); destructive feedback loops, by contrast, undermine positive relationships (introducing distrust, contempt, or other negative emotions into extant friendships) but ameliorate negative relationships (introducing sympathy, respect, or even admiration into extant enmities).

In sum, being a friend is not just causally but constitutively dependent on there being another person who has the same virtue. It is, second, not just causally but constitutively dependent on there being another person towards whom you harbor certain *de re* attitudes, and who reciprocates them. Third, it is not just casually but constitutively dependent on your thinking of yourself as someone's friend. Fourth, it is also not just causally but constitutively dependent on there being between you and your friend at least two orders of mutual knowledge of these attitudes. Fifth, it not just causally but constitutively dependent on you and your friend have first-, second-, and perhaps even third- and fourth-order emotions that include the other person in their content. Finally, it is not just causally but constitutively dependent on you and your friend sometimes knowing (and perhaps knowing that you know) that you are engaging in joint planning.

One might worry that these arguments press too hard on the relational aspects of friendship. Surely, one might think, I can be a friendly person even if everyone else is an asshole and either snubs or betrays my attempts at friendship. There is an important sense in which, even in such an unlucky social environment, I can still be a friendly person. This is a fair point, and one which should lead us to distinguish between the disposition or trait

of friendliness or agreeableness or gregariousness, which is arguably a monadic property of an individual, and the virtue of friendship, which clearly is not. One test that seems to do a good job of drawing this distinction is to ask whether the person in question is *friendly* or *a friend*. There is a double dissociation between these. Someone might be friendly but still not have any friends. Conversely, it's possible for someone to be dispositionally grumpy or unagreeable but nevertheless to be a friend, provided one's grumpiness or unagreeableness doesn't become so pronounced that it turns into outright misanthropy and make one unsuitable to be anyone's friend. A modicum of grouchiness can even be charming.[1]

3. Motivating the Friendship Model

In this paper, I explore the prospects for using the features of friendship identified above as a model for trust. This exploration is motivated on four independent grounds, which I discuss below.

3.1 The Historical Motive

Of the ten books of the *Nicomachean Ethics*, fully two are devoted to discussing friendship.[2] This is twice as much attention as justice receives, and as much as all of the other moral virtues combined. Yet contemporary neo-Aristotelian treatments of virtue rarely address friendship, and give it short shrift when they do. Annas (1993, pp. 249-260) devotes twelve of the five-hundred-plus pages of her book to friendship, and mentions it only twice in her (2011, pp. 76, 151) book. Geach (1977, p. 80) mentions it once. Hurka (2001, pp. 35-6, 200) mentions it in only a couple passages. Hursthouse (1999, p. 11) calls friendship an "awkward exception" because it is relational. MacIntyre (1981, pp. 123, 156-8) mentions friendship only twice. Russell (2009) barely engages friendship in his massive tome. Slote

[1] For an example, consider the case of a husband whose wife goes on anti-depressants to help her cope with her mother's recent death. The anti-depressants work too well, and she flips from being amusingly sarcastic to overbearingly cheerful. It can happen! (http://www.slate.com/articles/life/dear_prudence/2013/04/dear_prudence_my_wife_s_personality_has_changed_since_going_on_paxil.html)

[2] I should note that, though I draw here on what I take to be Aristotle's conception of friendship, I am not offering an interpretation of Aristotle.

(2001) only mentions friendship in the context of broader discussions of love, community, and achievement. Snow (2008), despite the fact that her book is titled '*Virtue as Social Intelligence'*, never once uses the word 'friendship'. Adams (2006, pp. 25-7, 69-92) and Roberts (2013, ch. 7) are the exceptions that makes the rule.

The main topic of discussion in the contemporary literature on friendship is the extent to which various moral theories induce "moral schizophrenia" by calling on us to be motivated by abstract principles – such as maximizing good outcomes or acting from duty – that seem incompatible with the warmth and intimacy of friendship (Stocker 1976). While this is an interesting issue, it is only tangentially related to the central questions of friendship: *What does it mean to be a friend? Is friendship a virtue? How is friendship related to trust, hope, and other attitudes? How is friendship related to more-commonly-discussed virtues, such as trustworthiness, generosity, and pride?* It would be surprising if this bias in the scholarship did not distort our understanding of virtue.

3.2 The Moral Psychological Motive

If there is a consensus in moral psychology, it's that virtue-concepts are "thick," in the sense that they refer to properties that are at once descriptive (and explanatory) and evaluative. To call someone a liar is to make an assertion about what sorts of behavior that person tends to engage in, but also to evaluate his behavior or behavioral tendencies by a normative standard. To think of someone as temperate is to attribute some behaviors or behavioral tendencies to her, but also to evaluate her behavior or behavioral tendencies by a normative standard. If you find out that someone about whom you have an otherwise good opinion tends to lie, you will feel considerable pressure to revise your good opinion or to find some excuses for his lying. If you find out that someone you would otherwise consider temperate has repeatedly ended up vomiting wine into a gutter, you will feel considerable pressure to revise your good opinion or to find some excuse for her excesses.

Friendships seem to have the same thick character as virtues. If you are my friend, I (and, for that matter, third parties) can form well-founded descriptive expectations about how you will behave, what you will think and feel, and how you will deliberate. If I am your friend, you (and, for that matter, third parties) can explain some of my behaviors, thoughts, feelings, and deliberative processes. In addition, if you are my friend, I (and third parties)

will tend to engage in characteristic evaluations. For instance, I will tend to give you the benefit of the evaluative doubt, and will – when I cannot find a way to do so – feel some pressure either to cut off our friendship or at least to scale it back. Moreover, realizing that you have inadvertently befriended a despicable person naturally engenders self-doubt: what sort of person am I, that I could befriend someone like this? What sort of person do I appear to be, that someone like this would want me for a friend? Additionally, those who think of you as a good person will, when they find that I am your friend, tend to form positive evaluations of me (if I am a stranger to them), or either positively revise their opinion of me (if they think well enough of you) or negatively revise their opinion of you (if they think ill enough of me).

In addition, it's generally recognized that virtues are threshold concepts: you can be honest enough to count as an honest person even if you've lied or cheated or stolen once in your life. You can be generous enough to count as a generous person even if in many (probably even most) cases in which you've had the opportunity to give something to someone who would enjoy or appreciate it, you did nothing. Honesty and generosity come in degrees. To count as honest or generous full stop, one needn't be at the furthest end of the spectrum: one simply needs to embody *enough* of the trait in question. Friendship also has this characteristic. I can be your friend even if we're not BFFs. Friendship comes in degrees. I can be better friends with X than with Y even though I'm friends with both X and Y. If, however, the quality of my relationship with X suffers too much, we cease to be friends at all. Where exactly to draw these lines is tricky business that I won't try to spell out in detail in this paper, but then, where exactly to draw the line between generosity and non-generosity is probably just as difficult.

Friendships thus seem to share with virtues both evaluative thickness and threshold instantiation properties; this is evidence that they belong in the same category. Exploring the features of one may illuminate the features of the other.

3.3 The Empirical Motive

As I have argued elsewhere (Alfano 2013, 2014), there are empirical grounds for doubting whether virtue as conceived in currently-dominant neo-Aristotelian theories is an achievable ideal. John Doris (1998, 2002), Gilbert Harman (1999), and I (2013) have

argued that most people's conduct does not seem to be structured by robust, global dispositions such as honesty – at least when they are tested in a decontextualized laboratory setting. Seemingly trivial and normatively irrelevant situational influences, such as mood modulators and ambient sensory stimuli, predict and explain people's cognitive, affective, evaluative, and behavioral responses as well as – and sometimes better than – personality variables. This is not the place to delve deeply into the dialectic between philosophical situationists and defenders of neo-Aristotelian ethics. Instead, I merely want to point out that there is suggestive empirical evidence – much of which I canvass in my (2013) book – for the phenomenon of *factitious* virtue. A factitious virtue simulates its neo-Aristotelian counterpart through the stabilizing influences of self-concept and social expectation-signaling. Someone may not be disposed to think, feel, and act as a generous person would think, feel, and act *except insofar* as she both thinks of herself as generous (self-concept) and knows both that others think of her as generous and that they know that she knows that they think of her as generous (social expectation-signaling). When this happens, she does not have the trait of generosity construed in neo-Aristotelian terms, but she does have factitious generosity.

Factitious generosity thus mirrors several of the more striking structural features of friendship. You cannot be my friend unless I think of you as a friend, and you know that I do, and I know that you know that I do. You cannot be factitiously generous unless I think of you as generous, and you know that I do, and I know that you know that I do. You cannot be a friend if you don't think of yourself as a friend. You can't be factitiously generous if you don't think of yourself as generous.[3] Thus, in addition to the historical and moral psychological motives for taking friendship seriously as a model for the virtues, we also have an empirical motive for exploring such a model. It may be possible to satisfy all of these motives by reconstructing other moral virtues on the model of friendship – as essentially and constitutively social.[4]

---

[3] Several other contributions to this volume, including those from Jacobs, Sifferd, Holroyd & Kelly, Webber, Athanassoulis, and Masala agree that character may be *causally* dependent on externalia. The novel contribution of this this chapter is argue that it may be *constitutively* dependent.

[4] All? Perhaps not humility and modesty, which seem to involve a paradox of self-reference insofar as it's hard, though maybe not impossible, to be humble and modest if

3.4 The Externalist Motive

The final motive for exploring the friendship model of virtue is the compelling evidence that has begun to pile up for the idea that many seemingly individual psychological phenomena are better understood as extending beyond the limits of the skin of the person to whom those phenomena are ordinarily attributed. In the 1970s, Kripke (1972) and Putnam (1975) popularized the idea that mental content is external, that the meaning and reference of some words is not determined solely by what's in the heads of people who use those words. In the 1980s, Nozick (1981) and Dretske (1981) introduced the notion that one's justification for a given belief might not be determined solely by what's in one's head. In the 1990s, Clark and Chalmers (1998) suggested that the mind itself might extend beyond the limits of the skin. According to their *parity principle*, if "a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in accepting as part of the cognitive process, then that part of the world is (for that time) part of the cognitive process." Importantly, they think that external phenomena can only be recruited in this way if they are reliably available, typically invoked, automatically endorsed, and easily accessible. I would suggest that these conditions should be relaxed slightly, such that the external phenomena be *at least as reliable* as the relevant internal phenomena, *at least as likely and quickly* to be endorsed, and *as easily accessible*. In a similar spirit, Adams (2006, pp. 138-43) argues that affiliations, such as being a Christian or a Communist, and social roles, such as being a (good) father or a (good) teacher, may be plausibly counted as moral virtues. Sneddon (2011) argues that processes of moral reasoning, responsibility attribution, moral judgment, and even action production are sometimes not just extended but *shared* by pairs or even larger groups of agents. Pritchard (2010) argues that cognitive abilities – which we might think of as intellectual virtues or parts thereof – may extend. Likewise, I have argued (Alfano forthcoming a) that the phenomenon of stereotype threat is evidence that cognitive processes extend.

The current proposal is that we should explore the extent to which this research program can be applied to virtue ethics, that is, the extent to which it makes sense to say that some

---

you also think of yourself as humble and modest. See Roberts & Wood (2007, ch. 9) for a discussion of this problem. I explore the possibility of such a paradox of self-reference with some of my colleagues in Alfano et al. (forthcoming).

or even all psychological dispositions that we might reasonably call moral virtues also extend beyond the limits of the skin of their possessors. The friendship model is a promising framework for doing so. At least when we are at our best, we try to live up to our friends' expectations; we are attuned to their reactive attitudes; we consider prospectively whether they would approve or disapprove of some course of action; we consult with them explicitly, implicitly, and imaginatively; and we revise our beliefs and values in light of their feedback. Friendship thus seems to involve the sort of functional integration that qualifies as an instance of extended character. Millgram (1987, p. 368; see also Cocking & Kennett 1998, p. 504) argues that, over the course of a friendship, each friend becomes causally, if not constitutively, responsible for the other's being who she is. Morton (2013) argues, in the same vein, that sometimes our behavioral dispositions are best governed by imagining the emotional reactions of those we love and respect to our thoughts, feelings, and plans.

Thus, we have four independent motivations – historical, moral psychological, empirical, and externalist – for exploring the friendship model of virtue. None of these is decisive, of course, nor is their conjunction. Nevertheless, I think that, together, they provide compelling reasons to give the model a chance. In the balance of this paper, that's what I do.

4. Trustworthiness on the Friendship Model

To begin fleshing out the friendship model, I want to concentrate on the virtue I take to be most amenable to assimilation: trustworthiness. Many philosophers have noticed that friendship often involves trust. Nehamas (2010, p. 238), for instance, points out that "friendship is immune, or at least resistant, to slander: we know our friends well and it takes much to undermine our faith in their goodness; [friends] trust one another." Likewise, Thomas (1987, p. 217) claims that one of the three most salient features of friendship is that "there is an enormous bond of mutual trust between" friends. Trustworthiness is a good prospect for assimilation to the friendship model of virtue.

Trustworthiness, as I describe it in this section, has many of the interesting structural features of friendship. It is not just causally but constitutively dependent on there being another person who has a congruent virtue (trustingness). Second, it is not just causally

but constitutively dependent on there being another person towards whom you harbor certain *de re* attitudes, and who reciprocates with their counterparts. Third, it is not just causally but constitutively dependent on your thinking of yourself as trustworthy. Fourth, it is also not just causally but constitutively dependent on there being between you and your trustor at least two orders of mutual knowledge of these attitudes. Finally, trustors and trustees are connected by characteristic first-, second-, and even third- and fourth-order emotions, such as the emotion of trust itself, as well as assurance, betrayal, outrage, repentance, forgiveness, and others.

4.1 The Cunning Mechanism

Consider the example, due to Pettit (1995), of a driver in an unfamiliar city who relies on a local bus driver to guide him to the city center. He notices that the bus displays a sign saying 'Downtown,' and so feels comfortable following the bus. However, since he realizes that the bus driver may find his consistently stopping when the bus stops disturbing, he pulls up beside the bus at some point to explain what he's doing. At this point, he isn't just passively relying on the bus driver. The bus driver now *knows* that he is being relied on, and the car driver knows that the bus driver knows this. Having put his trust explicitly in the bus driver, there exist at least two orders of mutual knowledge of this reliance between them. As Pettit (p. 205) puts it:

> The driver knows that I am relying on him and knows that I am aware that he knows that. Perhaps the reliance even becomes a matter of common knowledge, with each of us being aware of the reliance, each being aware of this awareness, each being aware of that higher-order awareness, and so on.

Just as, when we are friends, I wish you well, and you know it, and I know that you know it, so in such a case of interactive trust, I rely on you, and you know it, and I know that you know it.

There is a further point of analogy. When we are friends, I expect our relationship to count as a reason for action for you. Perhaps you wish me well for my own sake and in virtue of my good character, but our being friends provides you an additional reason to wish me well. Just so in the case of trustworthiness:

> I may expect that the driver will be positively moved by seeing that I have made myself vulnerable and will be motivated all the more strongly to do that which I am relying on them to do: will be motivated all the more strongly to prove reliable. (Pettit 1995, p. 206)

I take it that it is intuitively compelling that the bus driver would or at least might be thus motivated, but why? Why exactly would the bus driver sense an additional normative reason to prove reliable just because he is being explicitly trusted? Pettit's answer (pp. 214-5) to this question is that one of the more important rewards in moral psychology is esteem. People want money, power, sex, and stuff. They also want to be well-regarded – perhaps only instrumentally, but perhaps also intrinsically. When the car driver explicitly puts his trust in the bus driver, he manifests a form of a positive regard. This is, in itself, a reward to the bus driver, assuming that he cares about how she's perceived. And, since people are usually loss-averse (Tversky & Kahneman 1991), any value the bus driver attaches to the car driver's newly-acquired esteem will tend to be over-valued; he will attach more negative value to losing this esteem than he attached positive value to gaining it in the first place. Although Pettit does not make this further point about loss-aversion, it fits nicely with his view.

Finally, interactive trust of this kind essentially involves certain emotional dispositions, including higher-order emotional dispositions. At the first-order level, I will feel gratitude towards the bus driver for agreeing to lead me to the city center. He may feel anxiety on my behalf or pride that I chose to trust him rather than someone else. Should he prove reliable, I will experience another episode of gratitude, which may induce gratitude in him as well. Should he prove unreliable, I may experience anger. Should he deliberately lead me astray, I will probably experience betrayal. In response to the former, he may experience guilt; to the latter, malicious glee (if he endorses his betrayal) or surprise (if he didn't realize that I might react thusly).

4.2 The Self-Concept Mechanism

Along the same lines, we might point out that people also typically find self-esteem instrumentally and even intrinsically rewarding. When I signal that I think of you as trustworthy by explicitly putting my trust in you, I prompt you to revise your self-concept, to accept that you merit this trust. To the extent that you feel that you do, you will find your newly-revised self-concept rewarding. And, just as loss-aversion makes it more painful to lose others' high regard than it is pleasurable to gain it, so loss-aversion also makes it more painful to give up your own high self-regard than it is pleasurable to gain it. Hence, to the extent that you find it difficult to engage in the self-deceptive psychological acrobatics needed simultaneously to maintain your own self-esteem and to betray my trust, you will be motivated to prove trustworthy.

In addition, self-concept helps to set defaults for behavior, thought, and feeling. To the extent that I think of myself as honorable, I will be more inclined to try to avenge perceived offenses, think of others' actions as impinging positively or negatively on my honor, and feel offended or honored by others' actions and inactions. My self-concept thus constrains what I am inclined to notice, deliberate about, believe, feel, and do. Just as someone who thinks of herself as X's friend will perceive, think, feel, and act in different ways from someone who does not think of herself as X's friend, so someone who thinks of himself as patient will perceive, think, feel, and act in different ways from someone who does not think of himself as patient. Consider, for example, the following lyrics from Fountains of Wayne's song, "Michael and Heather":

> Michael and Heather on the shuttle bus
> Standing alongside the rest of us
> Michael says, "Heather, have you had enough?"
> Heather says, "Michael, you know that it's you I love."

Does Michael think of himself as Heather's lover? If he does, he'll probably construe her as answering the questions, "Have you had enough of this airport nonsense?" and responding somewhat playfully, "None of this nonsense really matters." But if he thinks of himself as merely pursuing Heather, it's likely that he'll notice a darker reading of her response. Perhaps she's not thinking about whether she's had enough of *waiting at the*

airport but of *their relationship*. Perhaps she's actively considering whether to break up with him, and is treating his question as an invitation to initiate the breakup. What we think of ourselves partially determines what even occurs to us as a possibility – a possible action, a possible interpretation, and so on. By influencing each other's self-concepts, then, we indirectly influence each other's conduct.

4.3 The Hopeful Mechanism

There are still further reasons to think that trustworthiness and trustingness can be assimilated to the friendship model. For, as McGeer points out, Pettit's explanation of how trust can be self-reinforcing is not the only possible rationalization of the relevant phenomena. As she explains, the mechanism Pettit identifies is unstable because it arguably would not survive its own elevation to mutual or common knowledge:

> trustees cannot know or suspect that they are only being trusted because the trustor is relying on the likelihood of their having a desire for good opinion; for then trustees will know or suspect that trustors do not really hold them in high regard (as actually possessing trust-attracting virtues), but only imagine them to be manipulable because they possess the less admirable trait of seeking others' good opinions. Hence, trustees will lose the incentive provided by a trustor's trust to act in a trust-responsive way. (2008, p. 252)

If the only reason I have to trust you is that I think that by putting my trust in you I can induce you to desire the continuance of my (signal of) high regard and the continuance of your own new-found high self-regard, then, were you to discover this reason, you might lose the very motive I aimed to induce. If you realize that my show of trust wasn't motivated by high regard but by an assumption about your desire for high regard, you may cease to value my show of trust. This does not mean that the mechanism Pettit identifies is never active, but it does suggest that there may be other, more stable mechanisms that produce similar results.

One is the mechanism of self-esteem and loss-aversion, which I sketched earlier. To my knowledge, it has not been identified previously. Even if you were to discover that my

expression of trust was an attempt at manipulation, if you had at that point already revised your self-concept, you would still presumably find it more aversive to go back to your old self-concept than you found it rewarding to revise in the first place.

Another mechanism, identified by McGeer, involves the attitude of hope. For her, hope essentially involves trusting beyond what the available evidence supports.[5] McGeer reminds her readers that human motivation is often complicated and confusing. Sometimes we don't know what we really desire, like, or love. Sometimes, we forget what we really value. In those cases, it's often helpful to refer to a normative lodestone, a model of conduct. She goes on:

> For help in this regard, we are sometimes encouraged to look outside ourselves for role models, finding in others' thoughts and actions laudable patterns on which to fashion our own. And this may serve us pretty well. However, something similar can occur, often more effectively, through the dynamic of hopeful scaffolding. Here we look outside ourselves once again; but instead of looking for laudable patterns in others' behaviour, what we find instead are laudable patterns that others see—or prospectively see—in our own. We see ourselves as we might be, and thereby become something like a role model for ourselves. The advantage in this is clear: Instead of thinking, 'I want to be like her,'—i.e., like someone else altogether—the galvanizing thought that drives us forward is seemingly more immediate and reachable: 'I want to be as she already sees me to be'. (2008, pp. 248-9)

This seems like a plausible explanation of at least some trustworthiness induced by acts of trust. It also fits nicely with existing discussions of friendship, according to which friendship is an especially salubrious context for acquiring self-knowledge. This illuminates the Aristotelian dictum that a friend is another self (see also Cooper 1977a, p. 300). It can sometimes be easier to know and understand oneself through the reflective mirror of what a friend sees in you than through introspection. And, if the extended character thesis is on the right track, there might be no fact of the matter concerning what your character is like

[5] Hope thus clashes, *prima facie,* with epistemic rationality. Stroud (2006) likewise argues that friendship clashes, *prima facie*, with epistemic rationality – another point of analogy.

until it is reflected back to you in the eyes of another person. Your trustworthiness might essentially involve your knowing that another person trusts you, or even that another person trusts you in a spirit of hope. The friendship model seems to satisfy all of the criteria that Clark and Chalmers (1998) identify as necessary for extended cognition: reliability, typicality, endorsement, and accessibility. Cues of trust tend to be reliably available – certainly as reliably available as other incitements to trustworthiness. Whether someone trusts someone else is typically invoked in predicting, explaining, and evaluating behavior. When someone with whom I have an ongoing relationship of trust expects me to do something, feel a certain way, or think something, I generally endorse this expectation automatically. Finally, the expectations and sentiments of my trustors tend to be easily accessible to me. I generally know what they want me to do.

Between McGeer's hope-based mechanism and my own self-concept-based mechanism, we may account for much of the factitious trustworthiness that Pettit's model doesn't cover. Trustworthiness and trustingness might then be said to constitute an interrelated and non-reducible dyad in much the same way that your being my friend and my being your friend constitute an interrelated and non-reducible dyad. Your being trustworthy would depend not just causally but constitutively (if only in part) on my trustingness, and my trustingness would depend not just causally but constitutively (if only in part) on your trustworthiness. Your trustworthiness would depend not just causally but constitutively on my harboring certain *de re* attitudes (of reliance, trust, and hope) towards you, and on your reciprocating with their counterparts (reliability, assurance, etc.). Your trustworthiness would depend not just causally but constitutively on your thinking of yourself as trustworthy (and hence, through loss-aversion, being unwilling to betray my trust). Your trustworthiness would depend not just causally but constitutively on there being between us at least two orders of mutual knowledge of these attitudes. Finally, your and my sentiments in a trusting relationship tend to be highly attuned, to the point that they will easily generate characteristic first-, second, and even third- and fourth-order emotions in appropriate conditions.

5. Conclusion

In this section, I briefly summarize the argument thus far, then offer a few remarks about my naturalistic methodology.

5.1 Summary of the Argument Thus Far

I've now argued both that we have reasons to take the friendship model seriously and that there are understandable naturalistic mechanisms through which it could work. If the considerations explored in section 2 are on the right track, then we have four main reasons to explore the friendship model. First, doing so reconnects virtue theory with its historical roots. Second, doing so helps explain why friendship, like other virtues, is a thick concept. Third, doing so may help to overcome or sidestep the situationist challenge. Finally, doing so may help to establish useful connections with externalist discussions in philosophy of mind. Naturally, all of these connections may also work in the other direction, helping virtue theorists export their insights and arguments to ancient philosophy, moral psychology, empirically-informed ethics, and externalist philosophy of mind.

Furthermore, if the explanations explored section 4 are on the right track, there are at least three mechanisms that, together, help explain how trustworthiness and trustingness are interdependent as the friendship model suggests. When the cunning mechanism is activated, one person's trustworthiness depends on another person's trustingness because the trustworthy person is motivated to prove himself worthy of the trust and esteem that have been directed his way. In this case, he wants to prove to the person who trusted him that he was and is worthy.[6] When the self-concept mechanism is activated, once again, one person's trustworthiness depends on another person's trustingness because the trustworthy person is motivated to prove himself worthy of the trust and esteem that have been directed his way. In this instance, though, he wants to prove *to himself* that he is worthy, in order to hold onto his positive self-concept.[7] Either way, esteem drives him forward. When the hopeful mechanism is activated, one person's trustworthiness depends on another person's trustingness because that trustingness gives him self-knowledge, opens up some actions as genuine possibilities, forecloses other actions as beyond the pale, and instills him with vicarious confidence. These are exactly the sorts of mechanisms that characterize friendship and that make it a paramount moral good.

5.2 Methodological Remark

[6] For more on this, see Wong (2006), especially chapter 4.

[7] For more on this, see Appiah (2011).

In this paper, I've described a model of trusting relationships. Is this model descriptive or normative? Does it characterize how trust *actually, typically* (or at least often) works, or does it characterize how trust *ought* to work? This question is prompted by the more general concern whether, in philosophical psychology, we should aim to elaborate descriptive or normative models. It might seem that I've given up on normativity altogether. One might be tempted to object, with Browning's Andrea del Sarto, that "a man's reach should exceed his grasp, / Or what's a heaven for?"

In response, I want to argue that ethics, like political philosophy, would benefit from a distinction between ideal and non-ideal theory (Rawls 1999). Moreover, I want to argue that we neglect non-ideal ethical theory to our own peril. One potentially fruitful way to proceed in virtue theory is to describe in rich detail how things actually, typically (or at least often) work, then use that description as an anchor for describing how they might work better. The ideal theorist wants to describe how things could be optimally. But there are reasons to prefer baby steps.

First, because ethics is a practical domain, ethical theory (including virtue theory) should at least sometimes be a useful guide to action. If you ask me the way to Larissa and I tell you that it's at 39°38'13" North by 22°25'13" East, I haven't given you an optimal answer. Similarly, if you ask me the way to Larissa and I tell you to go *that* way for an hour and then ask someone else, I haven't given you an optimal answer. But both answers are useful, and they're especially useful in tandem. Suppose for the sake of argument that the neo-Aristotelian picture of virtue is an adequate ideal theory. That doesn't mean that it's the only thing we need to know. Suppose for the sake of argument my friendship model of virtue suggests a few steps one could take towards being more virtuous. That doesn't mean it's the only thing we need to know.

Second, ethics – especially virtue ethics – might best be construed as essentially developmental. The point of virtue is not to achieve it and then rest content in your achievement. The point of virtue is to take another step away from where you started and another step towards a better way of living.

Third, as Morton (2012) convincingly argues in the context of intellectual rather than moral virtues, perfection can be the enemy of the good. Suppose that it would be better to achieve normative ideal X than to achieve normative ideal Y. But suppose that one's reach often does exceed one's grasp – that striving for X typically or even always yields at best an approximation of X, and that striving for Y typically or even always yields at best an approximation of Y. The following argument is clearly invalid: If X is better than Y, then an approximation of X is better than an approximation of Y. Now suppose, as I suspect many orthodox virtue theorists would say, that intrinsic trustworthiness is superior to factitious trustworthiness. In other words, suppose that trustworthiness whose vehicle is the trustworthy agent is morally superior to trustworthiness whose vehicle is that agent along with someone else who trusts her. It doesn't follow that an approximation of intrinsic trustworthiness is superior to an approximation of factitious trustworthiness. That depends on how well we approximate them. The jury is still out on that question. The right attitude to take, then, is not dismissiveness but curiosity. [8]

[8] With thanks for helpful discussion to Brian Robinson, Alex Madva, Asia Ferrin, John Cooper, Alexander Nehamas, Benjamin Morison, Daniel Wodak, Philip Pettit, Marcus Arvan, John Richardson, Anthony Carreras, Jon Webber, Dhananjay Jagannathan, David Morrow, David Wong, Alberto Masala, Owen Flanagan, Hallie Liberto, and Kate Manne.

Works Cited

Adams, R. (2006). *A Theory of Virtue*. Oxford: Oxford University Press.

Alfano, M. (forthcoming). *Nietzsche's Socio-Moral Psychology*. Cambridge UP.

Alfano, M. (forthcoming a). Stereotype threat and intellectual virtue. In Flanagan & Fairweather (eds.), *Naturalizing Epistemic Virtue*. Cambridge UP.

Alfano, M. (2014). What are the bearers of virtues? In H. Sarkissian & J. Wright (eds.), *Advances in Moral Psychology*, 73-90. New York: Continuum.

Alfano, M. (2013). *Character as Moral Fiction*. Cambridge University Press.

Alfano, M., Robinson, B., Stey, P., Christen, M., & Lapsley, D. (forthcoming). Intellectual humility: The elusive virtue. *The Journal of Positive Psychology*.

Annas, J. (1993). *The Morality of Happiness*. New York: Oxford University Press.

Annas, J. (2011). *Intelligent Virtue*. New York: Oxford University Press.

Appiah, K. A. (2011). *The Honor Code: How Moral Revolutions Happen*. Norton.

Aristotle. *Nicomachean Ethics*. In J. Barnes (ed.) and W. D. Ross & . O. Urmson (trans.), *the Complete Works of Aristotle, volume 2*, pp. 1729-1867. Princeton UP.

Clark, A. & Chalmers, D. (1998). The extended mind. *Analysis*, 58:1, 7-19.

Cocking, D. & Kennett, J. (1998). Friendship and the self. *Ethics*, 108:3, 502-527.

Cooper, J. (1977a). Friendship and the good in Aristotle. *The Philosophical Review*, 86:3, 290-315.

Cooper, J. (1977b). Aristotle on the forms of friendship. *The Review of Metaphysics*, 30:4, 619-648.

Doris, J. (1998). Persons, situations, and virtue ethics. *Nous*, 32:4, 504-540.

Doris, J. (2002). *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge: MA: MIT Press.

Geach, P. (1977). *The Virtues*. Cambridge: Cambridge University Press.

Harman, G. (1999). Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society*, New Series 119, 316-331.

Hurka, T. (2001). *Vice, Virtue, and Value*. Oxford: Oxford University Press.

Hursthouse, R. (1999). *On Virtue Ethics*. Oxford: Oxford University Press.

Jensen, A. & Moore, S. (1977). The effect of attribute statements on cooperativeness and competitiveness in school-age boys. *Child Development*, 48, 305-307.

Kripke, S. (1972). *Naming and Necessity*. Oxford: Blackwell.

Lewis, D. (2002). *Convention*. Oxford: Wiley-Blackwell.

MacIntyre, A. (1981). *After Virtue*. Notre Dame.

McGeer, V. (2008). Trust, hope, and empowerment. *Australasian Journal of Philosophy*, 86:2, 237-54.

Millgram, E. (1987). Aristotle on making other selves. *Canadian Journal of Philosophy*, 17:2, 361-376.

Morton, A. (2013). *Emotion and Imagination*. Polity.

Morton, A. (2012). *Bounded Thinking: Intellectual Virtues for Limited Agents*. Oxford UP.

Nehamas, A. (2010). Aristotelian *philia*, modern friendship? In B. Inwood (ed.), *Oxford Studies in Ancient Philosophy*, pp., 213-47. Oxford UP.

Nozick, R. (1981). *Philosophical Explanations*. Cambridge, MA: Belknap Press.

Pettit, P. (1995). The cunning of trust. *Philosophy and Public Affairs*, 24:3, 202-25.

Pritchard, D. (2010). Cognitive ability and the extended cognition thesis. *Synthese*, 175, 133-51.

Putnam, H. (1975). The meaning of meaning. *Philosophical Papers, Vol. 2: Mind, Language, and Reality*. Cambridge: Cambridge University Press.

Rawls, J. (1999). *A Theory of Justice*. Harvard UP.

Roberts, R. (2013). *Emotions in the Moral Life*. Oxford UP.

Roberts, R. & Wood, J. (2007). *Intellectual Virtues: An Essay in Regulative Epistemology*. Oxford UP.

Russell, D. (2009). *Practical Intelligence and the Virtues*. Oxford: Oxford University Press.

Slote, M. (2001). *Morals from Motives*. Oxford UP.

Sneddon, A. (2011). *Like-Minded: Externalism and Moral Psychology*. Cambridge, MA: MIT Press.

Snow, N. (2008). *Virtue as Social Intelligence: An Empirically Grounded Theory*. New York: Routledge.

Stocker, M. (1976). The schizophrenia of modern moral theories. *Journal of Philosophy*, 73.

Stroud, S. (2006). Partiality in friendship. *Ethics*, 116:3, 498-524.

Thomas, L. (1987). Friendship. *Synthese*, 72, 217-236.

Tversky, A. & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 106:4, 1039-61.

Wong, D. (2006). *Natural Moralities*. Oxford University Press.