This is a repository copy of *Norms of Trust*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/93988/

Version: Accepted Version

Book Section:
Faulkner, PR (2010) Norms of Trust. In: Haddock, A, Millar, A and Pritchard, D, (eds.) Social Epistemology. OUP Oxford . ISBN 0199577471

Reuse
Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

7

Norms of Trust

Paul Faulkner

1

Norms are instructions how to act. Their form is the imperative or hypothetical imperative: 'Do X', 'Don't do X', or 'if Y, then do X'. There are moral norms, epistemic norms, norms of practical rationality and social norms. In following norms we get to believe truths, be justified in our belief, act rationally, lead a virtuous life, and act in morally permissible ways. In following social norms we get to lead a life that is acceptable to a given society, a life that conforms to this socially established way of living. Social norms thereby differ from other norms in that their prescriptions are relative rather than universal. At the most trivial level these norms can be no more than a matter of etiquette: they are a matter of adopting the right register, wearing the right clothes, or, for instance, using the outer most knife and fork first. In these matters most would accept that when in Rome one should do as the Romans do. But this easy going relativism towards the prescriptions of social norms is not always so easy: social norms can be strongly felt. And nor are all social norms trivial: adopting the right register, for instance, can shade into treating someone the right way and thereby engage moral norms. Similarly, prescriptions of politeness can overlap with epistemic prescriptions. 'Believe people' might be a matter of politeness, but it could also be an epistemic norm if truth-telling were the norm. Is 'tell the truth' a social norm, and so the norm? Should we believe people? This paper

aims to argue that roughly speaking we have these social norms. It aims to offer an explanation of these social norms. And it aims to offer an account of how this bears on epistemological theories of testimony as a source of knowledge.

2

The attitudinal hallmark of social norms, Elster suggests, is that they are associated with, and sustained by, "feelings of embarrassment, anxiety, guilt and shame that a person suffers at the prospect of violating them, or at least at the prospect of being caught violating them. Social norms have a *grip on the mind* that is due to the strong emotions their violations can trigger."[1] Where the social norm concerns the behaviour of one party towards another, there are three attitudinal dimensions that could be individuated. Violation of the norm will provoke emotions of shame or guilt in the 'wrongdoer'; it will provoke the reactive attitude of resentment in the 'wronged'; and it will provoke punitive attitudes of disapproval or anger in third parties. These hallmark emotions are found in our attitudes towards truth-telling and believing others. This might be illustrated for truth-telling as follows. A native to this city you are approached by someone who is visibly a tourist and asked directions to the train station. Most would think it would be the wrong thing to do to misdirect the tourist, and that it would be shameful to misdirect the tourist 'for the fun of it'. A friend X has confided in you. The matter is of some delicacy. Another friend Y is curious about what is going on with X, you try to avoid the issue but Y is persistent and asks about the matter directly. With no room to manoeuvre you choose to maintain X's confidence and lie to Y all the while chaffing at being put in this position. These cases illustrate, at the very least, that in the context of being quizzed for information we feel that we *should* give the audience the information needed. Equally, we would be susceptible to guilt-like emotions were we to mislead the audience – even if

[1] Elster (1989: 100).

for good reason as in the case of the nosey friend. And we feel it would be appropriate for the misled audience to resent being misled. These hallmark emotions then suggest that we have some kind of norm of truth-telling. That there is a parallel norm of believing others is then illustrated by imagining things from the other side. Suppose that you look like the kind of person whose directions would be authoritative but after having given clear and confident directions to the train station you witness the tourist walk off in the opposing direction and promptly ask someone else for the same set of directions. This manifest distrust has something of an insult to it.[2] If humour didn't intervene, it would be liable to provoke something like resentment: what reason could he have had for not believing you? And one would expect the tourist to be embarrassed if he sees you watching him. This set of emotional responses equally suggests that we have something like a norm of believing others. So the first question is what is the content of these two norms?

The 'norm of truth-telling', I think, is less one of truth-telling, and more one of being cooperative in conversation. Suppose that as an audience you need to know whether *p* and engage a speaker in conversation with the purpose of finding this out. Ideally you want to engage with a speaker who will tell you that *p* if *p* and tell you that not-*p* if not-*p*. Ideally you want a speaker who makes her conversational contribution one that is true. But since it is not always plain what is true, the most one can really expect is for a speaker not to say what she believes is false or for which she lacks adequate evidence. But this is not all. Williams gives the example of being told "Someone's been opening your mail", when it is the speaker who has been doing so. This speaker has said something true, but what she has said is misleading because it implies the falsehood that *someone else* has been opening the mail. To communicate a truth, she'd have to say more: "Some has been opening your mail and that someone is me", (or "I've been opening

[2] A point observed by both Austin (1946) and Anscombe (1979).

your mail"). So if you want to know whether *p*, then, in addition to wanting a speaker to try and say what is true, you also want the speaker to be as informative as is required for our not being misled. Ideally the speaker's contribution would also be appropriately relevant and lucid. And if it is all of these things, the speaker's contribution has been guided by Grice's maxims of Quality, Quantity, Relation and Manner respectively. Given our conversational goal of learning whether *p*, a reply that is guided by these maxims is cooperative.

On the other side, the norm of believing others is less one of credulity and more the paired norm of presuming cooperation.[3] Grice's claim is that we can presume conversation to be cooperative; we can expect it to be guided by "the Cooperative Principle":

> We might then formulate a rough general principle which participants will be expected (ceteris paribus) to observe, namely: make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. One might label this the Cooperative Principle. (Grice 1967: 26)

For Grice, it is the presumption that participants are following this principle that makes a "talk exchange" a "conversation" rather than a "succession of disconnected remarks". This presumption, his point was to observe, can be needed to work out what a speaker means by what she says. For example, a speaker *S* states "You're a fine friend" on learning that her close friend *A* has divulged her secret to a business rival. In knowing that *S* knows of his actions, *A* knows that *S* believes that what she says is false. So *S* appears to be flouting the maxim of Quality and with it the Cooperative Principle. But they still seem to be having a conversation: *S* appears to be telling him something. So *A*

[3] Adler labels this norm the default rule: "one ought simply to accept a speaker's testimony unless one has special reason against doing so." (2002: 143). And he observes that the "default rule actually functions as a *presumption* that our informant's are being cooperative." (2002: 154).

must presume that *S* is following the Cooperative Principle and is telling him the opposite of what she says. There is little cooperation between *A* and *S* but *A*'s receiving the information from *S*'s telling that he is a poor friend requires the conversation be a cooperative endeavour. Now *not all* talk exchanges are cooperative: not all talk exchanges are *conversations* in Grice's sense. We might expect our interlocutors to cooperate in conversation, but they need not. This is illustrated by the mail case. The speaker *S* purportedly tells the audience *A* something. Since the conversation thereby has the seeming purpose of *S* giving *A* some information, *A* will expect *S*'s utterance to be such as is required, and to be true and appropriately informative. On the presumption that *S* is following the cooperative principle, *A* will thereby understand *S* to be telling him that someone else has been opening his mail. And there is room here for *S* to say truthfully that she didn't say *that*, and so has been misunderstood. She has been misunderstood because the presumption of cooperation is false; it turns out that they are not having a conversation in Grice's sense: *A* thinks that he is being told something when in fact he is being manipulated. But *A* was right to presume that they were having a conversation – *A* was right to presume the Cooperative Principle was being followed – not merely because this is how things seemed, but also because the Cooperative Principle is a normative principle amounting to the prescription that if you want to have a conversation which has a certain understood purpose, then you'd better make your conversational contribution such as is required by this accepted purpose. If the accepted purpose is the giving and receiving of information, then you'd better try and say what is true and be appropriately informative.

When a speaker satisfies these maxims, let me say that the speaker is *trustworthy*. Grice's proposal that participants in a talk exchange should follow and be presumed to follow the cooperative principle is then the proposal that we expect speakers to be trustworthy when the talk exchange has the accepted purpose of giving and receiving

information. The idea that conversation be seen as a cooperative endeavour thereby yields a pair of social norms. The prescription that speakers follow the Cooperative Principle and its maxims describes a *social norm of trustworthiness*. And the paired prescription that as audiences we presume this of speakers and act as if we believe that they are following the Cooperative Principle and its maxims describes a *social norm of trust*. Together this pair of norms describes a standard that we expect interlocutors to live up to when engaged in a certain practice: that of having a certain type of conversation. On this standard if another depends on you for information, then other things being equal you should try to say what is true and try to be appropriately informative; and if another purports to tell you that something is so, then other things being equal you should explain this in terms of their trying to be appropriately informative. The fact that the presumption that speakers follow the Cooperative Principle reveals what speakers mean by what they say so in a way that naturally describes how we understand others I then take to be good evidence that this pair of norms describes our conversational practices. Grice's definition of a conversation is not meant to be a mere term of art. This answers the first question of the content of the norms of truth-telling and belief: they are actually norms of trustworthiness and trust. The next question is what explains our having these social norms? What explains the fact that talk exchanges tend to have the civility of conversations?

3

An influential account of social norms is provided by David Lewis (1969). Lewis's starting point is the idea of a *coordination problem*. The resolution of these problems, Lewis argues, gives rise to conventions which then define social norms. We often need to coordinate our actions. A simple case: if we live in the suburbs on opposite sides of the city and decided to meet for a drink after work in the city we need to arrange a time and

a place. Supposing neither of us cares whether we meet at six or seven or at *The Lion* or *The Lamb*, then we have what Lewis defines as a coordination problem: we have a coincidence of interests in that of the four possible time and pub combinations it doesn't matter to either of us which pair is chosen but each likes one pair best given the other's choice. The problem is settling on a choice. If you choose seven at the *The Lamb* that is fine by me and best for me, and we could reach this arrangement by declaration and agreement. However, another possibility is coordination by precedence. If you get cut off just as you were about to make this suggestion, one arrangement may remain salient and this is a repeat of last week's meeting. If this arrangement then successfully repeats – we both turn up at the same time and place – it will establish an expectation of future conformity and we will have an embryonic convention: to meet at *The Lamb* at seven after work on Wednesdays. This gives Lewis's first approximation:

> A regularity R in the behaviour of members of a population P when they are agents in a recurrent situation S is a *convention* if and only if, in any instance of S among members of P,
>
> (1) everyone conforms to R;
>
> (2) everyone expects everyone else to conform to R;
>
> (3) everyone prefers to conform to R on condition that the others do, since S is a coordination problem and conformity to R is a proper coordination equilibrium in S. (Lewis 1969: 42)

This definition does not contain any normative terms, however social norms, Lewis suggests, can be defined as "regularities to which we believe one ought to conform" (1969: 97), and conventions are norms in this sense. This 'ought' Lewis explains in two ways. First, conformity to a convention is in everyone's interest: the regularity is a proper coordination equilibrium or a combination that each likes better than any other, given the others' choices. And, on Lewis's refinement, it is *common knowledge* that this is so

(1969: 52 ff). Consequently, everyone will recognise two reasons why one ought to conform: one ought to do what is in one's interests; and one ought to do what others expect one to do when this expectation is reasonable. And this expectation is reasonable given that conformity is to a regularity that everyone knows is in everyone's best interest. The 'ought' of social norms then carries the force of these two reasons on Lewis's account. It also has further force, which is detailed by Lewis's second explanation: the 'ought' of social norms is attached to feelings of approval and disapproval. If others are confronted by an action that fits with their reasonable expectations, then they are likely to approve; and if others confront an action that runs counter to these expectations, then they are likely to explain it discreditably. So a failure to conform to a convention elicits disapproval, and this constitutes a sanction. This makes conventions "by definition, a socially enforced norm: one is expected to conform, and failure to conform tends to provoke unfavorable responses from others" (Lewis 1969: 99). Sanctions offer additional motivation for compliance. So conventions are social norms because they have normative force: one has reasons to comply with them *and* there are sanctions on one's compliance. This explains why the regularities that constitute the norm persists. What explains there being regularity in the first place is that norms as conventions are proper coordination equilibria, and so in everyone's best interest.

4

In order to give this Lewisian explanation of the Cooperative Principle, and more specifically the norms of trust and trustworthiness, these norms must prescribe courses of action in a situation whose outcome is determined jointly by the actions of two or more. This they do. This situation – call it the *testimonial situation* – is a conversation whose ostensible purpose is the giving and receiving of information. The prescribed outcome is for the speaker to be trustworthy and so to try to tell the truth in an

informative way and for the audience to trust and so to act as if he believed the speaker is doing just this. In order for these norms to have the status of conventions, this outcome must be a proper coordination equilibria: it must be an outcome that each likes best given the other's choice. Now in any particular case this could be true. But conventions and norms must hold independently of the particularities of any given case. And it is easy to imagine cases where this outcome is not a proper coordination equilibria. Holding the audience's trust constant, a speaker might have preferred the outcome where he was untrustworthy. Suppose the speaker has just made a kill in a hunt and would rather keep this kill for himself and his family. In this case, if another hunter asks him whether he had any luck hunting, he might well tell the other hunter the truth, but he would prefer the outcome where there is trust but he keeps the fact that he has made a kill concealed. So in this particular case the prescribed outcome is not a coordination equilibria. Moreover, if the testimonial situation is considered in abstract there is no reason to think that this case is peculiar. The recurrent situation is that of a conversation whose ostensible purpose is the giving and receiving of information, wherein the audience ostensibly needs to know whether *p*, and the speaker ostensibly tells the audience what he needs to know. The choices for the audience are to trust or not – to accept what the speaker tells him on the presumption that speaker is trustworthy or not; and for the speaker to be trustworthy or not – to try to tell the audience the truth informatively or not. Given the assumptions that a false belief would leave the audience in a worse position than ignorance and that the audience's interest is what it appears to be and that is being informed, the best outcome for the audience is to trust when the speaker is trustworthy and not to trust when the speaker is not trustworthy. Call the first outcome the *cooperative outcome*. Whilst, in any particular case, it may lie in the speaker's interest to tell the truth what explains this lie of interest will be the more basic interest of influencing the audience. A conversation whose ostensible purpose is the giving and

receiving of information will be of interest to a speaker because it is a way of getting an audience to believe something and is thereby a way of exerting an influence on the audience. And we have a basic interest in exerting an influence on others. Telling the truth could further this interest but only telling the truth expediently, or telling the truth conditional on one's other interests. Being bound to give an audience the information sought – being trustworthy – would not be in a speaker's interest. So whilst the cooperative outcome would be the best case outcome for an audience, the commitment of trustworthiness would not be best for a speaker. The best outcome for a speaker would be to receive an audience's trust yet have the liberty to tell the truth or not given the shape of interest in the particular case. However, telling the truth merely when it suits one is a way of being untrustworthy. So the best outcome for a speaker would be the trust and untrustworthiness combination. Since this is the worst case scenario for an audience, the testimonial situation has a structure of pay-offs that resembles the prisoner's dilemma.[4] As such the testimonial situation could be said to present a *problem of trust*: the rational thing for an audience to do seems to be to not trust another for information in the first place.[5] Since the testimonial situation is thereby not one "in which coincidence of interest predominates", it is not a coordination problem in Lewis's sense. So it seems that Lewis's explanation of social norms cannot be given for the norms of trust and trustworthiness.

5

[4] This point is observed by Adler: "Testimonial situations allow modelling as repeated Prisoner's Dilemmas at least, presumably, in their origins (when testimony is more functional and local, and so defection more readily detected)." (2002: 155). And Pettit claims that the norm of "Telling the truth reliably rather than expediently, randomly, or whatever" is "equivalent to cooperating in a many-party prisoner's dilemma"(1990: 735).

[5] See Faulkner (2010).

Nevertheless, Lewis's explanatory strategy can still be pursued. This strategy is to explain the cooperative outcome in terms of the interests of the parties producing it. Ultimately what explains a convention is the fact that it is a proper coordination equilibrium. The problem is that in prisoner's dilemma type cases non-cooperation is the dominant option, in that it is the rational thing to do whatever the other party does. However, this all changes given a couple of assumptions about the recurrent situation that the presents the dilemma. The first assumption is that individuals have a sufficiently large chance of meeting again so that they care about future interactions, and the number of these future interactions is indefinite. I might depend on you for information on this occasion but you will depend on me on some future occasion and there is the mutual expectation that we will find ourselves in an indefinite number of these testimonial situations. As such there is no final interaction where the trusted party might hope to secure extra benefit by behaving untrustworthily. The second assumption is that each party can make it clear to the other that they will behave in a tit-for-tat manner, which is cooperating at the start but responding to any non-cooperation with retaliation. If the hunter concealed the fact that he made a kill on this occasion and was found out, he would lose the benefit of sharing on another occasion. So whilst non-cooperation could potentially pay dividends on any given occasion, this would be offset by future losses. Given these two assumptions, the cooperative outcome is a coordination equilibrium: no one benefits from unilateral non-cooperation. And it is a proper coordination equilibrium: there is no better strategy. That the cooperative outcome then comes to describe a convention and so be prescribed as a social norm follows on Lewis's account once it is added that everyone expects and prefers conformity. This expectation is delivered by the second assumption: it is clear to all that each is behaving in a tit-for-tat manner and so will conform by default. And the preference for conformity follows given that another's non-cooperation results in loss of the benefits of the cooperative outcome for two rounds,

this and the next when retaliation is due. So it is possible to give Lewis's explanation of how the cooperative outcome becomes a matter of convention, where this explanation appeals to little more than rational self-interest.[6]

Philip Pettit (1990: 735) develops a particular implementation of this strategy for explaining norms of cooperation. He starts with three criticisms. First, what this strategy explains is not a norm prescribing a certain behaviour but one prescribing that one behave in this way in a tit-for-tat manner. So it does not offer an explanation of the norm of trustworthiness, or telling the truth reliably, so much as one of the norm of truth-telling in a tit-for-tat way. However, Pettit does not regard this as a significant matter because these norms should be extensionally equivalent in that general tit-for-tat truth-telling would result in the same behaviour as reliable truth-telling. Second, this explanation only works for certain norms. It only works for what Pettit calls "type B prisoner's dilemmas":

> In a type B dilemma, defection by even a single individual plunges at least one cooperator, and perhaps many more, below the baseline of universal defection. In a type A dilemma this is not so and, at limit, the lone defector may have only an imperceptible negative effect on cooperators. (Pettit 1990: 737).

The problem for this explanatory strategy is then that lots of many-party prisoner's dilemmas are type A. However, this is not a problem in this case: the problem of trust is type B because defection by a speaker in the testimonial situation will plunge an audience below the baseline; that is to say, it is worse for the audience to have a false belief than remain ignorant. The real problem, third, is that behaving in a tit-for-tat way requires that people be willing retaliate. And if the norm being explained is a norm of trustworthiness as opposed to tit-for-tat trustworthiness, then being trustworthy in a tit-for-tat way

[6] It is developed in detail in Axelrod (1984). See also Blais (1987).

requires "that people will break norms punitively in order to punish those who break them for convenience. … And this disposition is not genuinely manifested among those who honor norms" (Pettit 2990: 737). Pettit's key observation then follows this third criticism and pursues Lewis's thought that sanction can come by way of others' disapproval. It is that whilst people might not be given to retaliation – such behaviour is costly and risks an escalation of retaliation rather than a return to cooperation – people do retaliate after a fashion: "a violater can be punished – or of course a conformer rewarded – by the attitudes of others" (Pettit 1990: 739). And "once these approbative costs and benefits are put into the equation", Pettit hypothesises, "we can see our way to explaining why the emergence and persistence of otherwise puzzling norms maybe unsurprising" (1990: 742). That is, once these costs and benefits are put into the equation we can make good the tit-for-tat implementation of Lewis, (except we now understand 'tatting' as holding a disapproving attitude). What is assumed here is that people care about others' approval and disapproval. If it is also assumed that these feelings are automatic and costless, then one gets sanctions for free and one has an explanation of compliance with the norm: the sanction which ensures compliance is not the threat of retaliation but the fear of disapproval and desire for approval. So we can regard trustworthiness as the best preference given trust, and the cooperative outcome as a coordination equilibrium given the approval it elicits. Thus starting with Lewis's suggestion that conventions are norms because they specify regularities which everyone recognises that one ought to conform to. And making it explicit that the feeling *that one ought to conform* underwrites approval and disapproval and plays a causal sustaining role gives Pettit's Lewisian account of social norms:

> A regularity, R, in the behaviour of members of a population, P, when they are agents in a recurrent situation, S, is a *norm* if and only if, in any instance S among members of P,

(1) nearly everyone conforms to R;

(2) nearly everyone approves of nearly anyone else's conforming and disapproves of nearly anyone else's deviating; and

(3) the fact that nearly everyone approves and disapproves on this pattern helps to ensure that nearly everyone conforms. (Pettit 1990: 731).

In short, norms are regularities in behaviour that we approve of where this approval and corresponding disapproval enforce the regularity. The approval and disapproval are rationally intelligible because the norm specifies the cooperative outcome in some situation of collective action. And this outcome can be seen to be a proper coordination equilibrium once our concern for approval is factored into the equation.

6

Pettit distinguishes two different ways of explaining norms of cooperation. 'Behavioural explanations' show how certain patterns of behaviour are a matter of self-interest. The tit-for-tat explanation of cooperation is behavioural; it shows how cooperation can "emerge in a world of egoists without central authority" (Axelrod 1984: 3). Pettit contrasts his 'attitude-based explanation' which shows "first why certain attitudes of approval are intelligible … [before] then showing how they might generate the patterns of behaviour required for norms" (Pettit 1990: 733). Now suppose that agents *S* and *A* are in a joint action situation T and there is a norm stating that *S* should ϕ in T. Pettit's account as to why disapproval of breach of a norm like this is intelligible is that the norm prescribes a cooperative outcome that is a proper coordination equilibrium. So were *S* not to ϕ, *A* would disapprove of *S*'s failure to ϕ because it is to his, *A*'s, detriment, and *A* thinks that *S* has a reason to ϕ. Now this does render *A*'s attitude of disapproval intelligible after a fashion: it renders it intelligible from the perspective of what lies in *A*'s rational interest. However, this is not intelligibility from *A*'s perspective in that it does

not detail *A*'s reason for his attitude of disapproval. The judgement of wrongness made when a norm is breached does not consist of a judgement of consequence: *A*'s disapproval does not stem from the loss that will follow from *S* not ϕ-ing. Rather, *A*'s judgement is no more than that *S* should have ϕ-ed, or should have ϕ-ed given that he was in situation T. Consider the question, what exactly do people disapprove of when they disapprove of others breaking a norm? Or, what makes disapproval intelligible to those making a disapproving judgement? These questions will be answered by appeal to the norm: the norm that one should ϕ in T describes a standard of conduct that is appealed to when making an approving or disapproving judgement. What this means is that one cannot see the attitudes of approval and disapproval as intelligible independently of acceptance of the norm that articulates these judgements and that is appealed to in order to justify them. Consequently, these judgements do not have the kind of independent intelligibility necessary for Pettit's 'attitude-based' explanation. That is, *unless* intelligibility is really just a matter of rational self-interest. But then patterns of behaviour are ultimately approved of because they are in everyone's best interest, and the distinction Pettit draws between styles of explanation is not substantive. But this is just to say something that Pettit (1990: 725) acknowledges, which is that his account is game-theoretical.

On Pettit's definition of social norms, general attitudes of approval and disapproval "help to ensure" conformity to the regularity in behaviour that is the outward part of the social norm. This is undoubtedly true, but the question is what is the right reading of "helps to ensure". The right reading for Pettit's 'attitude-based' explanation is a strong one like 'is the cause of'. The idea is that conformity is explained once people's preferences are suitably adjusted to take into consideration the fact that we care about others' opinions of us. The problem for this explanation starts when we take seriously the question, what exactly do people disapprove of when they disapprove of

others' breaking a norm? In terms of the schematic scenario just described, what *A* disapproves of when *A* disapproves of *S* not ϕ-ing in T is just the fact that *S* didn't ϕ when *S* should have done. The norm itself articulates *A*'s reason for disapproval. Social norms are thereby held as standards of behaviour that justify both attitudes of approval and disapproval. To use a sociological term of art, people *internalise* social norms, which is to say that social norms describe patterns of behaviour that people are motivated to follow for no other reason than that these patterns of behaviour are valued in themselves or held as ultimate ends. Let me say describe this kind of motivation as *intrinsic*.[7] The idea that we internalise social norms is the idea that we are instrinsically motivated to behave in the way the norms prescribe. This aspect of social norms, the fact as Elster says that they "have a *strong grip on the mind*", is not represented by game theory, which is not concerned with intelligibility from the inside or intelligibility in terms of the norm. So returning to the issue of *A*'s disapproval of *S* not ϕ-ing, this disapproval is the kind that is meant to explain, on Pettit's account, general conformity with the norm to ϕ in action situation T. But if *A*'s disapproval stems from the fact that *A* has internalised this norm then it is this fact and not others' disapproval that should explain *A*'s compliance. But what goes for *A* goes for *S*, these are just agent role placeholders. This is to say that whilst others' disapproval can cause conformity, the reason for others' disapproval will also centrally be their reason for conformity. Returning to Pettit's criticisms of the tit-for-tat behavioural explanation of social norms, these criticisms can now been seen to be more substantial than Pettit acknowledged. If peoples' reason for conformity is their having internalised the norm they conform to, then we have an explanation as to why people tend not to break the norm punitively. To break the norm punitively is still to break the norm and so fall short of the standard of behaviour prescribed. And if the norm itself is peoples' reason for conformity, then it matters whether the norm is to ϕ or

[7] I follow the terminology of Sripada and Stich (2006).

to ϕ in a tit-for-tat way. Even if these two norms are extensionally equivalent, they give different reasons for action. So the tit-for-tat explanatory strategy is limited from the start.

The idea that social norms articulate reasons for action and figure in justifications of attitudes of approval and disapproval allows for a very simple explanation as to why people conform to social norms. We obey social norms because these norms describe what we believe that we ought to do. We obey social norms because we have internalised them. This is more properly an 'attitude-based' explanation of social norms, and can be labelled the 'social' account (since talk of internalising social norms is commonplace in the social sciences). However, whilst it provides a very simple answer to the question of why people conform, the social account raises to two further questions. First, we might have a rudimentary explanation of why people comply with social norms, but can this be filled out? Second, game theoretical or Lewisian accounts of social norms make good sense of why norms exist in that they provide an explanation not merely of compliance but also of the norms themselves. So one might also hope for an answer to this question: why do particulars norm exist? In the next section I consider this question, and its particular form: why is it that we have social norms of trust and trustworthiness?

7

In *Truth and Truthfulness* Bernard Williams offers an imaginary genealogical account of what he calls the *virtues of truth*: *Accuracy* and *Sincerity*. These are the dispositions to care about the truth of one's beliefs, and to come out with what one believes. Williams's genealogy offers an explanation of *our valuing* these dispositions or virtues of truth. It is because we value these dispositions that we try to get things right in belief and utterance. However, our valuing Sincerity is not a matter of the crude prescription: 'always tell the truth'. The mail opener, referred to above, does not manifest the disposition of Sincerity

in Williams's sense, even though she is sincere in believing what she says. When the conversation is one of giving and receiving information, Sincerity is a matter of being appropriately informative: it is a matter following the cooperative principle and its maxims, which define the social norm of trustworthiness. So Williams's genealogical justification of Sincerity offers a way of explaining why it is that we have this social norm, and the paired social norm of trust. Moreover, Williams's genealogical justification focuses on the joint action situation that is the testimonial situation which presents the problem of trust.[8]

Williams's genealogy starts by imagining a State of Nature consisting of a primitive social group with limited technology and no writing. Although primitive, this social group is imagined to be a real society whose members have projects and interests, and are related to one another in various ways and via various roles. As with any society, the society imagined in the State of Nature will involve cooperative engagements which demand information be communicated between individuals. Given that an individual can only be at one place at one time, individuals will often gain what Williams (2002: 42) calls *purely positional advantage*; that is, by virtue of their location at a time, one individual can come to possess information that another individual needs. It follows that even in the State of Nature, thus minimally characterised, Accuracy and Sincerity are desirable from the social point of view; they will be socially valued because pooled information is a social good and necessary for many cooperative endeavours. However, possessing the disposition of Sincerity need not always be in an individual's best interest. Williams gives the example of the hunter who has found prey that he would rather keep for himself and his family. This raises the problem that:

[8] I discuss Williams's genealogy in more detail in Faulkner (2007).

> The value that attaches to any given person's having this disposition [Sincerity] seems, so far as we have gone, largely a value for other people. It may obviously be useful for an individual to have the benefits of other people's correct information, and not useful to him that they should benefit of his. So this is a classic example of the "free-rider" situation. (Williams 2002: 58).

The problem is that the collective valuing of Sincerity does not itself give an individual a reason to value Sincerity, or be sincere. Whilst it is always in an audience's interest to be informed, sincerity needn't best serve a speaker's interest and as audiences we know that this is the case. The problem that Williams thus identifies is the problem of trust; and since the State of Nature represents a basic society, the possibility that a conversation as to the facts could be stymied by this problem shows "that no society can get by … with a purely instrumental conception of the values of truth" (Williams 2002: 59).

What any society requires is that individuals have internalised Sincerity as a disposition, where this is to say that individuals are motivated to act in a sincere way simply by the description of this way of acting as sincere. Where this is true, Sincerity will have *intrinsic value* (or intrinsic value in the society). Something's having intrinsic value, Williams then goes on two argue, can be understood in terms of the satisfaction of two conditions. For something, X say, to have intrinsic value in a society: first X must be "necessary, or nearly necessary for basic human purposes and needs"; and second X must "make sense to them [the society members] from the inside, so to speak" (Williams 2002: 92). The first of these desiderata is established by the imagined genealogy. If Sincerity is not given intrinsic value, then any conversation that purports to be one of giving and receiving information will generate the problem of trust. This threatens to stymie both the conversation and any further cooperation. However, we do cooperate in conversations as to the facts. We tell one another what we know and we have a way of

life wherein testimony is a source of knowledge. So whatever needs to be in place to avoid the problem of trust must be in place and that is that we intrinsically value Sincerity. We must be motivated to be Sincere as an end in itself. The second desiderata is then giving an account of how this motivation is made sense of.

Since any value is made sense of through its connection to further values, how a society gives intrinsic value to Sincerity can be philosophically unearthed through conceptual analysis. What conceptual analysis shows about *our* social history is that *we* understand Sincerity through its relation to trust and our valuing trustworthy behaviour. *Sincerity is trustworthiness in speech*. This is more than the avoidance of lying; our being trustworthy can require our lying. Equally, it is not simply the disposition to say what one believes; one can implicate falsehoods by saying what one believes and so be untrustworthy by doing this. The mail opener implicates that someone else has been opening your mail. What Grice's discussion of implicature then shows is that:

> Implicature do not presuppose language as simply a practice involving semantic and syntactic rules, together with the norm that certain kinds of utterances are taken to be true; they look to the use of language under favourable social conditions which enable it to be indeed co-operative. They are *conversational* implicatures, but not everyone who is talking with someone else is engaged, in the required sense, in a conversation. What is required for that to be so are certain understood levels of trust. (Williams 2002: 100).

We have achieved these levels of trust because we intrinsically value trustworthiness in speech, which is the disposition of Sincerity. And this is just to say, I suggest, that the norm of trustworthiness, which is the prescription that speakers follow the cooperative principle and its maxims, is internalised as a social norm. In learning to have conversations one learns this norm, and the presumption that things are as the norm prescribes then allows us to uncover implicatures or what people mean by what

they say. Since we can tell people what we know by implication as much as by bald statement, our norm of trustworthiness is then necessary for testimony being the source of knowledge that is. The genealogical justification that Williams offers in *Truth and Truthfulness* for our having the disposition of Sincerity can then be presented as a genealogical justification of the norm of trustworthiness, and with it the paired norm of trust. This addresses the challenge of explaining why we have these particular norms.

8

The problem, as Williams is aware, is that the claim that X has intrinsic value faces a dilemma. Left like this it is mysterious as a claim. Why should the description of an act as X be a motivation to act this way? But if the mystery is explicated, then the account of the value threatens to become reductive with X being merely instrumental valuable. Williams's two condition account is meant to address this dilemma; genealogy is meant to achieve "explanation without reduction" (2002: 90). Now a similar problem faces social accounts of social norms. The explanation of behaviour that states that people act a certain way because they have internalised a certain norm is not fully satisfactory issuing in the question: but why should they be motivated to act in this way? This is the first challenge noted in section six, and here game theoretical, or Lewisian, accounts of social norms seem to be genuinely explanatory since they have self-interest as the final appeal, which does not seem to just raise further questions. The problem with this, I argued, is that if social norms are rationalised in terms of rational self-interest, then what is left out is the sense that we find in acting in the way that the norm prescribes. What is left out is our understanding of the value that motivates our action. However, to make good the claim that this genuinely impoverishes these explanations, what is needed is a fuller account of *how* our understanding of value motivates action. Moreover, the need to meet this challenge is sharp in this particular case because, on the face of it at least, we

do find trust problematic. That this is so can be easily reinforced by considering a testimonial situation but stripping away all the factors that could be used in a game-theoretical solution to the problem of trust. One must imagine that there is no sanction on untrustworthy behaviour and that the speaker does not care for the good opinion of others, or at least that the audience cannot have the assurance of believing either of these things. That is, one must imagine a case where an audience is engaged in a conversation as to the facts with a speaker whose particular motivations and preferences the audience is ignorant of. In this case, the worry that the speaker will not tell the truth, or will only do so if it suits them is a natural worry. Thus the idea that it is reasonable to trust "if we know absolutely nothing about someone", Williams describes as simply "a bad piece of advice" (2002: 111). He then adds that

> [i]t may be said that a hearer never has a reason for believing that P which lies just in the fact that a given speaker has told him that P. He has to believe also that the speaker (on such matters, and so on) is a reliable informant. (Williams 2002: 77-8).

So our intrinsically valuing Sincerity is *not sufficient*, according to Williams, to ground reasonable acceptance testimony as to the facts. What is also needed is the belief that a bit of testimony is reliable or that a speaker is manifesting the disposition of Accuracy. Now I think that this is the wrong way to go, and the wrong way for Williams to go: the attitude of trust can suffice for reasonable acceptance of testimony. And I think that this is what is delivered by considering how our valuation of Sincerity 'makes sense to us from the inside'. However, leaving this argument until the next section, the point to be made here is that the temptation to require more than trust for reasonable belief is an expression of finding the problem of trust genuinely problematic. So if this problem of trust is meant to be solved by our giving intrinsic value to Sincerity, then a fuller statement is needed as to how this locus of value motivates our acting in certain ways.

This is the first challenge facing social accounts: filling out how the internalisation of social norms explains conformity with them.

What is needed, the case of the norms of trust, is recognition of how trust figures in our explanations and justification of action. These explanations and justifications are rather straightforward. Asked why we took the risk of depending on someone we often answer simply that we trusted them, and asked why we put ourselves out to do something we answer that someone trusted us to do this thing. Suppose that one person *A* trusted another *S* to do something, to ϕ. And suppose that *A* trusted *S* to ϕ in the thick sense that *A* depended on *S* ϕ-ing and expected this to be at least part of *S*'s reason for ϕ-ing.[9] Such an attitude of trust is quite common and might be found in the coordination problem described in section three. In this case, *A* trusts *S* to turn up at the *The Lamb* at eight and *A* trusts *S* to do this in the sense that *A* thinks that at least part of *S*'s reason for turning up at this pub at this time is the fact that *A* depends on *S*'s doing so. In this case if *S* were asked why he was going to this pub at this hour he might reply 'to meet *A*' or '*A*'s waiting for me' and if pushed to take another course of action, *S* could emphasise the reason this gives by making it explicit: '*A* trusts me to turn up'.[10] So we can use the fact that another has trusted us to do something to explain why we did this thing. And we can use the fact that we trust someone to explain why we showed a willingness to depend on them in certain ways: asked why he was bothering to get to the pub on time, *A* might reply that he trusts *S* to show up at this time. So the attitude of trust, in this thick sense, figures in justificatory explanations of action. This, I suggest, is what the claim about intrinsic value amounts to: we credit these kinds of ways of making sense of things. Since we use the attitude of trust to explain and justify acts of trust and

[9] I tried to work out a definition of this thick sense of trust in Faulkner (2007).

[10] This explication is closer to the surface when more is at stake. Maybe it is a first date and *A* and *S* are enamoured, or *A* is a potential informant and *S* is trying to gain his confidence in a political climate where both have much to lose.

trustworthiness, our having the dispositions that follow from internalising norms of trust and trustworthiness 'makes sense to us from the inside'. However, to say that we use the terms of these norms to explain and justify, is to say that we are motivated to act in the same terms. This is what the idea of internalisation delivers: the prescription of the norm captures the way the subject thinks about the action prescribed. The idea that we have internalised the norms of trust and trustworthiness then offers a genuinely explanatory account of action because if it is true that we have these dispositions to trust and be trustworthy, then the prescriptions these norms make will outline good descriptions of our reasons for acting.

9

Williams's official solution to the problem of trust found in the State of Nature is that this problem is resolved by finding someway to give Sincerity intrinsic value. We then give Sincerity intrinsic value by taking it to be a form of trustworthiness and valuing trust and relations structured by trust. Now if this claim about how we give Sincerity intrinsic value is understood as a claim about our finding certain descriptions and justifications of action persuasive, then the shape of the solution to the problem of trust Williams's genealogy offers is revealed.

The problem of trust presupposes an account of the kinds of reasons people have for acting. On this account, action is explained in terms of the agents beliefs and desires. Epistemic rationality demands that an audience desires to believe the truth and avoid falsehood and so has a preference for ignorance over error. Then absent any belief about a speaker's motivations, or any grounds for predicting the probable truth of utterance, and the result can only be that it is not reasonable for the audience to accept what he is told. Game theoretical solutions to this problem add grounds for belief. A concern for others good opinion in Pettit's case. With the problem situation then reconfigured, a

coordination equilibrium is found and the norms of trust and trustworthiness emerge. Williams recognises the extent of the problem, offers the basis of a solution in terms of giving intrinsic value to Sincerity, but then feels compelled to add that some belief about the truth of utterance is still necessary. However, his idea that Sincerity be intrinsically valued suggests an alternative solution. On this solution, what goes wrong with the problem of trust, why trust is seen as problematic, is that the only explanation of acceptance allowed is one that proceeds in terms of an audience's belief and desires. However, we can act out of trust: our trusting a speaker for the truth can give us sufficient reason to accept what the speaker tells us. An audience can explain why he accepted what a speaker told him, and so detail his motivations and justify his acceptance, by saying that he trusted the speaker for the truth. However, an explanation of action that is couched in terms of trust cannot be translated into one couched in terms of belief and desire. This is because the attitude of trusting someone to do something involves placing a expectation on that person: that they will act in certain way and for a certain reason. And this *expectation of* them is not the *expectation that* something will happen. The difference between these kinds of expectation is marked by the fact that when we expect things of people we are susceptible to various reactive attitudes if they do not act as we expect. So given a conversation as to the facts and an audience who trusted the speaker for the truth and was misled, this audience will be liable to resent the speaker's actions. The susceptibility to such a feeling of resentment defines the expectation as one that is placed on a person or held of them and distinguishes it from the expectation or belief that something will happen. And this feeling of resentment involves commitment to the norm of trustworthiness as an objective standard: any resentment felt will not be mollified by the knowledge that the trusted individual had no inclination to tell the truth because what is felt is that the trusted individual *did have* such a reason and *should have* acted on this reason. This is the reason described by the norm of

trustworthiness, which is meant to prescribe behaviour irrespective of subjective motivation or personal interest. So the norms of trust and trustworthiness define standards of behaviour within which explanations in terms of trust make sense and are accordingly good explanations. This is what is missed when trust is seen as problematic.

Where norms of trust and trustworthiness are internalised, the social background will be one of "certain understood levels of trust". It will be such that if an audience *A* trusts a speaker *S* for information, this will give *S* a reason to tell *A* what he needs to know; and if *S* tells *A* something, this gives *A* reason to accept what *S* tells him. The idea that conversation can be structured by presumptions of trust then allows for the following explanation of what goes on, or should go on, in a conversation as to the facts. A speaker's reason for telling an audience what he does – the explanation of the speaker's testimony – will be the speaker's perception that the audience depends on him for this information. In this case, if the speaker *S* tells the audience *A* that *p*, it will be because *S* believes himself to know that *p*, and assumes responsibility for letting *A* know that *p* in the following sense: *S* takes it on himself to tell *A* that *p* if and only if he, *S*, knows that *p*. In this way *S* is trustworthy. And in recognising that the speaker *S* intends that he *A* come to believe that *p* and trusting *S*, *A* will then take *S*'s telling him that *p* as something like a promise that *p* is true. Testimony then functions to transmit knowledge from speaker *S* to audience *A* because it transmits the responsibility for justifying belief from audience *A* to speaker *S*. This explanation is offered by Richard Moran (2006) and following his lead call it the *assurance* explanation. On this explanation *A* doesn't need the belief that *S* is reliable to have a reason to accept what *S* tells him. Rather, *A*'s trusting *S* gives him a reason through delivering the presumption that *S* is trustworthy. Moreover, for *A* to seek further reason to believe that *p* would be to reject *S*'s assurance that *p* is true. This would be as likely to provoke *S*'s resentment as straight disbelief since, Moran argues, it amounts to a refusal to accept the *S*'s assumption of responsibility in telling *A*

that *p* (see Moran 2006: 301). And this feeling of resentment, I suggest, is parallel to that a misled audience would experience in that it equally involves commitment to a social norm as an objective standard, in this case the norm of trust. Supposing *A* did seek the support that *S* was reliable prior to belief, *S*'s resentment at not being believed would not be appeased by the knowledge that *A* had no inclination to belief because what is felt is that *A* did have a reason for belief, given by *S*'s telling, and should have believed for this reason. So violations of either norm of trust – disbelieving a speaker or misleading an audience – will engender resentment and other punitive attitudes. So let me drop the demand for a belief about Accuracy or reliability from Williams's genealogy: it is better to see the problem of trust confronted in the State of Nature as resolved by the establishment of levels of trust that allow the giving and receipt of testimony as assurance.

How does the existence of social norms of trust and trustworthiness bear on epistemological theories of testimony? It is now possible to give a brief answer to this question. One implication is that non-reductive theories are correct to describe our attitude towards what others tell us as trusting. However, non-reductive theories, I suggest, are wrong then to hypothesise that we are default justified in trust; the norm of trust is a social norm and not a general or universal epistemological principle. Whilst every society confronts the problem of trust, since it is confronted in the State of Nature, securing the necessary motivations through a valuation of trust is but one solution to this problem. Other norms are possible. Another implication is that insofar as these social norms do operate in our society, there should be plenty of scope to give a reductive theory of testimony: we should have good evidence that tellings will prove generally reliable even if evidence of particular reliability is hard to come by.[11] However, the possibility of this defence of reductive theory should not constitute a justification of this

[11] See Adler (2002: Ch. 5).

position because the norm of trustworthiness is associated firstly with testimony being a source of knowledge that can be explained in assurance terms. And this, I think, is the central implication. Our having the social norms of trust and trustworthiness is a function of our having a way of life wherein we have conversations as to the facts and tell one another what we know. The epistemology of testimony cannot be synonymous with the epistemology of tellings: there are too many messy and varied cases, which determine that the assurance can only be part of the story. But if we have these norms of trust and trustworthiness, then the assurance explanation of how another's telling can put us in a position to know something must be an essential part of our epistemological story.[12]

Faulkner, Paul. 2007. "A Genealogy of Trust". *Episteme* 4 (3):305-321.

———. 2010. Lies and The Problem of Trust.

[12] This paper was written whilst in receipt of AHRC funded research leave. My thanks to Adrian Haddock and an anonymous referee.