

On Trust and Philosophy

by Tom Bailey

Trust is as elusive in philosophy as it can be in practice. Philosophers often simply ignore or presuppose it, and when they do consider it, they often struggle to explain it or confuse it with other things. Nonetheless, by considering some major philosophers' thoughts on trust and related matters, we can reveal certain important features of it, and see why it might be so elusive, in both philosophy and practice.

The ring of invisibility

In *The Republic*, Plato recounts a dialogue between Socrates and Glaucon, Plato's older brother. In it, Glaucon argues that only the fear of detection and punishment prevents a human being from breaking the law and doing evil for the sake of his own self-interest. Glaucon thinks that this natural fact is demonstrated by the shepherd Gyges, who found a gold ring which made him invisible whenever he twisted it on his finger. (According to the story, he found the ring on a corpse in a hollow bronze horse, which was revealed when an earthquake opened up the ground beneath his flock.) On realising the ring's power, Gyges used it to seduce the queen, murder the king, and take the throne. Glaucon's claim, then, is that every one of us, however law-abiding and good we might seem, would do as Gyges did, or something else in our self-interest, if we could avoid detection and punishment. And, Glaucon claims, we would be right to do so, since each human being's only interest is their own self-interest, and we have no interest in justice and morality for their own sakes.

This sorry picture of human nature has been accepted by many philosophers since Glaucon, and, indeed, by many other people as well. It raises the following crucial question: when and why should we trust others, if we think that only the fear of detection and punishment prevents them from harming and stealing from us? Glaucon's answer is that we should trust others only if we are confident that they fear detection and punishment sufficiently to dissuade them from harming or stealing from us. Thus I should trust the doctor to prescribe me appropriate treatment – and not, for example, make me the unknowing guinea pig for a new, untested medicine – only if I think that she is sufficiently afraid of being found out and struck off if she does not.

This seems reasonable, and reveals an important feature of trust: when we trust others, we are confidently relying on them to take care of something which we care about, but which they could harm or steal if they wished. When we trust, then, we make ourselves vulnerable. But we do so in the confidence that the trusted will not exploit this vulnerability, and generally in the confidence that the trusted will actively take care of what we make vulnerable. This vulnerability and care can concern something tangible, such as when I trust my friend with my bike, or something less tangible, such as when I trust a stranger to be honest when I ask him the time.

Machiavellian implications

But Glaucon's answer to the question of trust also has a truly disturbing implication. This is that, if I am *not* confident that others are sufficiently afraid of detection and punishment, I should expect them to try to harm or steal from me for their own self-interest. I should therefore be prepared to defend myself and, indeed, to pre-empt their attacks with attacks of my own. Thus if I think that the doctor is so set on using me as a guinea pig that she is prepared to risk detection and punishment, I should consider self-defence classes, a stronger lock on my door, and the price of hit men. She, on the other hand, should expect me to consider such things, and so should take measures of her own to resist and overcome them. We are likely to quickly reach a dangerous, even catastrophic, state of mutual distrust. Glaucon's argument thus seems a recipe for disaster.

Glaucon, however, is confident that detection and punishment in Athens are sufficient to dissuade his fellow Athenians from crime, and he does not consider the disturbing implication of his argument. This implication was brought home forcefully to Niccolò Machiavelli, however. He had been second chancellor of the Florentine republic, but when the Medici family took power through a coup d'état, they accused him of plotting against them and subjected him to an excruciating torture called the '*strappado*' (the 'torn'). Recovering outside the city, he wrote *The Prince*, a book of advice for new authoritarian rulers such as the Medici. (He even proposed to send them a copy, in the hope of being offered a job.) In it, he writes, 'One can make this generalisation about men: they are ungrateful, fickle, liars, and deceivers, they shun danger and are greedy for profit; while you treat them well, they are yours...but when you are in danger they turn away'. Machiavelli concludes from this not just that the Medici should be careful whom they trust. Rather, Machiavelli proposes that they take the brutal Cesare Borgia as their example and be prepared to be cruel, murderous, dishonourable, deceptive, and miserly whenever necessary to maintain their power. 'I draw up an original set of rules', Machiavelli writes nonchalantly.

Glaucon's sorry picture of human nature seems to leave us in a sorry state. If we accept that human beings' only reason not to harm or steal from each other in the pursuit of their self-interest is fear of detection and punishment, then it follows that when I am not confident that others have enough such fear, I should be prepared to pre-empt their attacks with attacks of my own. Since they also know that I will be prepared to do this, they should take measures to resist and overcome my attacks, and our distrust and attacks will spiral, ending only with the victory of the most brutal and cunning.

Trust between enemies?

But before we ditch our qualms and adapt ourselves to the possibility of a world of mutual distrust, cunning, and brutality, we should first consider another, much more attractive response to Machiavelli's conclusion. For, enlightened about human nature by the example of Gyges and disturbed by its Machiavellian possibilities, it would surely be better if human beings could agree a truce, as a guard against the escalation of mutual distrust and attack.

Thomas Hobbes recognises the attraction of this response. His experience of the collapse of the English state into civil war had made him well aware of the Machiavellian possibilities of human nature, and in his *Leviathan*, he considers what human life would be like without a state. Without a state, we would live in a 'state of nature', with no authority to tell us what to do, and no agencies to detect and punish us if we do not do it. This again raises the crucial question of trust: in a state of nature, could we trust others not to harm or steal from us? And if we could not, could we avoid the dangerous escalation of distrust and attack by agreeing a truce?

In considering human life in a state of nature, Hobbes understands human nature in essentially the same way as do Glaucon and Machiavelli. Hobbes assumes, firstly, that a human being is moved only by his own 'passions', his particular desires for, and aversions to, particular things. Secondly, Hobbes assumes that no human being is strong enough to be entirely secure from harm by others. (He calls this our natural 'equality'!) It follows from this that we do not curb our desire for something just because someone else has it. Thirdly, Hobbes assumes that the things we want are generally either scarce (so that we cannot *all* get what we want) or relative (so that my getting more of a thing effectively means that you have less of it). For example, food might be scarce, while power can be relative. It follows from this that in pursuing the things we want, we must view each other with distrust, as enemies. And the best way to prevent others from getting the things I want is, of course, to attack them before they attack me. As Hobbes puts it, 'there is no way for any man to secure himself so reasonable as anticipation, that is, by force or wiles to master the persons of all men he can'. From this, mutual suspicions and attacks will spiral, and Hobbes reaches his famous conclusion: in a state of nature, there would be 'war...of every man against every man', and life would be 'solitary, poor, nasty, brutish, and short'.

Hobbes thus draws the same conclusion from Glaucon's picture of human nature as Machiavelli does: that if the fear of detection and punishment is not sufficient to dissuade people from harming and stealing from me, I should be prepared to attack them before they attack me. However, Machiavelli simply sings the praises of those most successful in their attacks on others, while Hobbes sees this war as a 'miserable condition', and understands that we would wish to avoid it. In particular, he recognises that we might wish to agree a truce amongst ourselves, an agreement to restrain the pursuit of our self-interest when necessary to avoid war.

This wish reveals a second important feature of trust: that trust is a means of making our social life simpler and safer, and of making possible cooperative activities which each of us could not undertake alone. Indeed, trust is required for many cooperative activities which seem to make human life both liveable and worth living, such as friendship and love, the growing of food, and the raising of children.

Wishful thinking

However, Hobbes argues that for anyone actually to act on the wish to agree a truce in a state of nature would be disastrous for them, and therefore foolish. Consider, first, a very modest agreement: we agree not to attack each other during a specific period – say, tomorrow. If we all keep to this agreement, our lives will be a little less solitary, poor, nasty, brutish, and short than they would otherwise have been. But consider how I reason when deciding whether to keep to the agreement. If I think others will not keep to it, then it would be foolish for me to keep to it, since I would thereby make myself vulnerable to attack from them tomorrow, and lose valuable attacking time of my own. On the other hand, if I think that others will keep to the agreement, then it will be better for me to exploit this by attacking them, when their guard is down, than for me to keep to the agreement too. Thus it is always better for me not to keep to the agreement, whether or not others will. Since they will also reason like this, we will all behave as if we had not made the agreement at all, and we will all remain in our ‘miserable condition’. As Hobbes puts it, ‘covenants [i.e., agreements] without the sword are but words, and of no strength to secure a man at all’.

Perhaps it is not surprising that we are tempted to break such a modest agreement for the sake of our self-interest, and one might think that a less modest agreement would be more robust. Imagine, then, that we agree not to attack each other for a week, and threaten to give up the agreement and return to attacking each other if any one of us breaks it during the week. Then it seems that each of us will have an interest in keeping to the agreement, since keeping to it on one day ensures that we will benefit from it on the remaining days of the week. Assuming that we are sufficiently concerned with this future benefit, then, it seems that each of us should keep to this less modest agreement.

But consider how I reason on the last day of the week. Then, there are no benefits from later days to consider, and so I will reason exactly as I did about the modest, one-day agreement. Others will also reason like this, and so we will all be prepared to attack each other on the last day, despite our agreement. But now consider how knowing this affects my reasoning on the penultimate day. Knowing that we will not keep to the agreement on the last day, I also have no reason to keep to the agreement on the penultimate day. Others will also have no reason to keep to it, and we will all not keep to the agreement on the penultimate day either. Knowing this, in turn, ensures that we will not keep to the agreement on the day before the penultimate day, which ensures that we will not keep to it on the day before that, and so on, until the whole week of proposed peace unravels.

The obvious way to prevent this unravelling is to make an even less modest agreement: namely, an agreement not to attack each other *indefinitely*, again with the threat that we will all return to attacking each other if any one of us breaks the agreement. In this case, there would always be future benefits to consider, but there would be no specific last day from which our destructive backwards reasoning could begin. If successful, this agreement would also establish a lasting, possibly everlasting peace, rather than just a day or a week of it.

But Hobbes doubts that even such an agreement would work. His reason for this is simply that not every human being acts rationally all of the time. We often reason badly, fail to consider the future, or are carried away by our feelings. In particular, Hobbes notes that some of us pursue things obsessively, beyond their actual use to us. He also notes that we often respond excessively when we think that others are not treating us with the respect we think we deserve. Thus, even when it is in the interest of each of us to keep to the agreement, some of us may fail to recognise this. And even those rational enough to recognise it may not be confident enough in others' rationality to keep to it themselves. Given the prevalence of irrationality among human beings, the uncertainty of knowing who might act irrationally and when, and the huge risks involved in keeping to the agreement, Hobbes concludes that even those rational enough to wish that the agreement be kept would be foolish to keep to it. As he puts it, 'it is a precept, or general rule, of reason that every man ought to endeavour peace, *as far as he has hope of attaining it*, and when he cannot attain it, that he may seek and use all helps and advantages of war'.

This is a profoundly depressing conclusion. It means that if each human being acts rationally in his own interests, and does not have sufficient fear of detection and punishment, he must pass up certain crucial opportunities to cooperate with others. Furthermore, he must do this even when he knows that, if others are rational, they will do the same and the outcome will be worse for everyone than if they had cooperated. Rationality thus demands that he make himself an exception to such cooperation in the hope that others are not rational enough to do the same, so that he may exploit their gullibility. In other words, it demands that a human being be Machiavellian. And this applies as much to friendship and love, the growing of food, and the raising of children as it does to truces in a state of nature. It is conclusions like this that lead some philosophers to give up their faith in rationality, since it seems to demand that human life should become neither liveable nor worth living.

A more humane human nature

In desperation, or perhaps irritation, one might respond by suggesting that Glaucon, Machiavelli, and Hobbes have all simply misunderstood human nature. For, one might insist, Gyges is not a good example of human nature, since he manifests nothing of our natural concern for others. David Hume, an Enlightenment philosopher of much more optimistic and genial temperament than Machiavelli and Hobbes, suggests this. He recognises that human beings naturally care for their loved ones and sympathise with others' feelings, including those of complete strangers. In his *Treatise of Human Nature*, he writes that sympathy makes human beings 'mirrors' of each other, and that this give them 'a remarkable desire of company, which associates them together, without any advantages they can ever propose to reap from their union'. Love and sympathy of this kind would ensure that there would be more trust, and therefore more cooperation and peace, in a state of nature, or between princes, than Hobbes and Machiavelli claim. For example, I may be able to trust the members of my family not to attack me in a state of nature, simply because I know that they love me.

But unfortunately, even among human beings who love and sympathise with others, there is still much scope for distrust and war. Firstly, there is the sad fact that human beings' care for their loved ones makes them badly disposed to enemies of their loved

ones. Think of the Fowlers and the Mitchells in *Eastenders*, or the Capulets and the Montagues in *Romeo and Juliet*. Hume rightly accepts this. He also recognises that, although sympathy for others can make us more impartial, neither it nor love necessarily overcomes our more egoistic interests. If Hume had seen *Eastenders*, he would not have been surprised to see Phil Mitchell having an affair with his brother's wife, despite the bond between the brothers. In a state of nature, such trumping of love or sympathy by self-interest might be particularly crucial: for example, you might sympathise with my hunger, or even love me, but not enough to give me your food. Finally, note that even a saintly human being, whose sympathy for others provides him with his most important interests, must compete with the less saintly (and, perhaps, other saints with different sympathies) for the things he cares about. This is what makes *Superman* films (barely) watchable.

However, one might still try to resist the Machiavellian conclusion. One might, for example, put one's faith in education and civilization to improve and spread our sympathy for others, and thus reduce the likelihood of distrust and war. This is Hume's hope, and that of many other Enlightenment figures. Many of them, like John Locke, Immanuel Kant and Karl Marx, even go so far as to presuppose a shared sense of morality, which might be cultivated to overcome the partiality of our self-interest, love, and sympathy. (Socrates responds to Glaucon's argument by making a similar claim.) But if we had sufficient sympathy for others, or really shared a sufficiently strong sense of morality, we would not seem to need to trust each other at all. Even assuming that such improvements are possible, then, putting our faith in them does not give us much guidance for living and trusting in our present, vulnerable condition, in which we must live until such improvements are made.

Alternatively, of course, we might follow Glaucon, Hobbes, and numerous Home Secretaries in turning to detection and punishment to dissuade us from harming and stealing from each other. The possibility of being found out by the police and punished with a prison sentence, for instance, might dissuade some potential wrongdoers from crime. Those particularly afraid of their fellows might even think that a very coercive state is a price worth paying for security. But detection and punishment can be effective only if the vast majority of human beings do not even think of harming or stealing from each other, and if we already have a certain trust in the agencies of detection and punishment themselves. To emphasise detection and punishment therefore seems, at best, to be a very limited solution, and, at worst, to miss the point entirely.

Our last hope of avoiding the Machiavellian conclusion, then, would seem to lie in human irrationality. Like those philosophers who have given up their faith in rationality, we could simply hope that human irrationality will somehow keep our most essential cooperative activities afloat. But this is a largely blind and decidedly risky strategy, and, indeed, seems paradoxical: how can we rationally persuade ourselves to be irrational? This strategy should therefore be attempted only if all else fails.

Rethinking

As at the end of every bad film, however, there is perhaps one last chance to avoid disaster. And it is our natural love and sympathy, and our sense of morality which point us in the right direction. These nobler parts of human nature suggest that when we trust others, we are confidently relying on their good disposition towards us – we are relying on their love or sympathy for us or their sense of morality, for instance, rather than on their egoistic interests, habits, or irrationalities. Thus trust is a special kind of reliance, reliance on others' good disposition towards us. In contrast, if I expect my friend not to steal my bike just because I have asked him to leave a deposit, then I may be relying on him not to steal it, but I am not trusting him. (Nor is he likely to remain my friend for long!) Similarly, if I rely on others not to attack me in a state of nature just because I believe that it is in their self-interest not to break our agreement and that they are rational enough to recognise this, again I may be relying on them, but I am not trusting them.

This implies that Glaucon, Machiavelli, and Hobbes cannot even conceive of genuine trust, since their picture of human nature does not allow for it, and that we need to move beyond their picture if we are to explain genuine trust. But we also need to recognise that trust cannot rest only on love, sympathy, or the sense of morality. The insufficiency of love and sympathy is particularly clear. For we can rely on others' love or sympathy without necessarily trusting them as well, and we can trust people who we do not believe to love or sympathise with us at all. I rely on my dotting grandma to make me shortbread every time I see her, but it would be odd to think that I *trust* her to do so. Rather, I just rely on her love for me and for cooking shortbread. We can therefore rely on others' love or sympathy without necessarily having to trust them as well. (To avoid offending my grandma, I should say that, of course, I do trust her as well.) We can also trust others without believing that they love or sympathise with us at all. I might find out that my grandma really despises me, but I could still trust her to make me shortbread every time I see her. This is particularly clear when we trust in institutions, officials, and professionals. For I need not think that the doctor loves or sympathises with me in order to trust her not to use me as a guinea pig for untested medicines. Even more clearly, it would be absurd to believe that those involved in the testing of medicines are reliable just because they love me or sympathise with patients in general. But I can trust them nonetheless. Indeed, we can say exactly the same thing about their sense of morality, and even their fear of detection and punishment.

Nor do we necessarily trust others just because we know that they have been reliable in the past. I have a great deal of evidence of my grandma's making me shortbread whenever I see her, but I do not therefore have to trust, or even rely on, her to do so; and I can trust the people who test medicines without having any hard or conclusive evidence about their reliability in doing so. Also, the capacity to forgive those we trust for unreliability, and their capacity to respond when we encourage them to be more reliable, can be crucial to the cultivation and maintenance of trust. When I trust, then, I am not simply making a judgement about the past reliability of the trusted, although of course I may take this into account.

Thoughts like these suggest that there is more to trust than even an extended picture of human nature, such as that offered by Hume, Locke, Kant, or Marx, would allow. For

the ‘good disposition’ of the trusted, on which we rely when we trust them, cannot consist only of their past reliability, their fear of detection and punishment, their love or sympathy, or their sense of morality. Trusting others might therefore seem to rest, ultimately, on an irreducible *feeling* about their good disposition, or even on a ‘leap of faith’, much like its Christian counterpart.

But one need not go this far. Below I will briefly suggest one way in which we can account for our belief in others’ good disposition towards us, and therefore rationally trust them, without reducing trust to reliance on any of the features of human nature considered so far. I will leave you to judge whether you think this way of understanding trust is plausible.

Taking responsibility

This way emphasises that human life is primarily social. This means that each human being must consider how others behave, and how they will respond to his own behaviour, in deciding how to act himself. Even Gyges and Borgia had to do this, in order to achieve their dastardly ends. Thus the possibility of relying on each other to behave and respond in predictable, manageable ways is particularly valuable for human beings. Now, such reliance can be ensured by the detection and punishment which Glaucon and Hobbes emphasise, the love and sympathy which Hume emphasises, or the sense of morality which Locke, Kant and Marx emphasise. But the heart of trust, as another way of ensuring such reliance, lies elsewhere. For my reliance on others can be ensured simply by their taking responsibility for how their behaviour will influence my decisions about how to act in a particular regard. For example, they can take responsibility for my health, my security, or my bike, and so take responsibility for ensuring that I can rely on them in making my decisions about my health, security, or bike. This taking of responsibility, rather than love, sympathy, or a sense of morality, is the ‘good disposition’, or ‘trustworthiness’, on which I rely in trusting another. Indeed, if I believe that the other appears ‘trustworthy’ *only* because it coincides with his own interests, or even his love, sympathy, or sense of morality, I cannot believe that he really is trustworthy and so cannot trust him. I can then rely on him only in the common sense, by relying on detection and punishment, or his love, sympathy, or sense of morality. I cannot genuinely claim to trust him if I believe that I can rely on him *only* by resorting to such things.

Such taking of responsibility is part of being a friend, a lover, or a spouse, and a particularly important part of being a professional, an official, or a politician. If I trust the doctor to prescribe me appropriate treatment, I rely on her because I believe that she has taken responsibility for her role in my decisions about my health. Indeed, I may even allow her to effectively make these decisions for me. Similarly, I may rely on a policeman or policewoman, a judge, or a politician because I believe that they have taken responsibility for using the coercive powers of the state in certain legitimate ways and for certain legitimate purposes. I may thus rely on them when I make decisions about my safety or my property, for example. And even in more personal relationships, when I can rely more on another’s love for me, I still cannot trust them as a friend, lover, or spouse unless I also believe that they have taken responsibility for the particular, intimate role which they play in my life.

We must often leave exactly how others may fulfil such responsibilities relatively indeterminate, just because we often lack expertise in the area concerned, and are unable to predict contingencies. In trusting, we therefore allow the trusted some discretion. But this does not give them *carte blanche* to do as they wish. Their taking responsibility implies that they cannot intentionally lead us to rely on them in ways they cannot or will not satisfy, since this would conflict with our basic reason for trusting them. They must therefore be at least competent and honest. Nor can they simply 'take responsibility' for something which we could not want them to take responsibility for. (Imagine a thief who claimed that he was just 'taking responsibility' for my bike!) And, although one cannot genuinely trust others if one resorts *only* to reliance on detection, punishment, love, sympathy, or a sense of morality, one can certainly make *some* use of such resorts without necessarily failing to trust. Making judgements about such matters again requires discretion, however, if we are to avoid replacing genuine trust with common reliance.

Conclusions

If we stick to a picture of human beings as moved only by self-interest, love, sympathy, or their sense of morality, then, the rationality of trust will remain obscure. Many of our most valuable cooperative activities will seem to be irrational, and will seem to persist only through blind habit or hope. Our friendships and our visits to the doctor will continue to be haunted by the Machiavellian conclusion: that if we are not confident that others are moved by self-interest, love, sympathy, or morality not to harm or steal from us, we should attack them before they attack us. If we act on this conclusion, mutual distrust and attack will spiral and we will soon find ourselves in a decidedly 'miserable condition'. But if, out of habit or hope, we do not act on this conclusion, we will be blindly putting our faith in irrationality to keep our cooperation afloat, and we will struggle to cultivate and maintain trust, because we will not fully understand it. We are thus likely to do it more harm than good.

Recognising that human beings may take responsibility for how their behaviour influences others' decisions, however, offers us a way of explaining how trust can be rational. It thus also offers us a way of beginning to understand how trust can be genuinely cultivated and maintained. This does not mean that trust will necessarily be any less elusive in practice, particularly among those who mistake common reliance for genuine trust or believe in the common picture of human nature. But given the potentially disastrous consequences of such misunderstandings, the importance of making trust a little less elusive in philosophy should not be underestimated.

Some accessible introductions to the philosophy of trust

For the basic issues and positions regarding trust, see the very clear account in Karen Jones's entry on 'Trust' in *The Routledge Encyclopedia of Philosophy* (Routledge, 1998). There is a shorter version of this in *The Concise Routledge Encyclopedia of Philosophy* (Routledge, 1999). For more extended discussion, Annette Baier's essays, 'Trust and Antitrust' and 'Trust and Its Vulnerabilities', are very interesting. They are both reprinted in her book *Moral Prejudices: Essays on Ethics* (Harvard University Press, 1994). Martin Hollis's book *Trust Within Reason* (Cambridge University Press, 1998) provides an extended discussion of the Machiavellian conclusion, along with an interesting suggestion for how it might be avoided.

Editions quoted in this essay

Thomas Hobbes, *Leviathan, with Selected Variants From the Latin Edition of 1668*, edited by Edwin Curley (Hackett, 1994).

David Hume, *A Treatise of Human Nature*, edited by L.A. Selby-Bigge, revised by P.H. Nidditch (second edition, Oxford University Press, 1978).

Niccolò Machiavelli, *The Prince*, translated by George Bull (revised edition, Penguin, 1999).