Value-Mistaken and Virtue-Mistaken Norms

Philip Pettit

Many norms appear to arise, or at least to stabilize and fixate, as a result of an error on people's parts as to the attitudes of others. I think that this is a phenomenon of more than marginal interest and my aim here is to put it in the limelight. In doing so I draw in part on work done elsewhere.[1]

My discussion will be in five sections. I look in turn at the definition of norms; at how, under this definition, a norm can be based on an error shared among those who observe it; at the psychological plausibility of such an error; at the sorts of norms that are likely to be supported, for good or ill, in this way; and at the lessons for institutional design.

1 The Definition of Norms

The word 'norm', as used in characterization of society, has two more or less obvious connotations. First, anything that deserves to be described as a norm of a society – an actual, not a would-be, norm – has to be a regularity that prevails there: a pattern of behavior that characterizes the society, marking it off from actual or potential competitors. But, second, the regularity in question cannot be a matter of indifference amongst the people who sustain it. In order to have the status of a norm, a social regularity has to be seen, however tacitly, as something with a certain sort of normative claim on people's allegiance: as something that, for whatever reason, is appropriate in relevant contexts; it has to attract approval or the breach of the regularity disapproval.

But these two connotations do not exhaust the associations of the word 'norm', as that is used in social contexts. Suppose that a regularity was behaviorally and attitudinally supported among the members of a certain society but that there was no connection between the attitudinal approval and the behavioral compliance; suppose, in other words, that the attitudinal support was epiphenomenal in relation to the behavioral compliance. In that case the regularity would certainly be something that people had reason to welcome and celebrate, like an aspect of their biology;

[1] Pettit (1990); Brennan and Pettit (1993, 2000, 2004).

J. Kühnelt (ed.), *Political Legitimization without Morality?*

but it would scarcely count as a regularity that they treated as normative. This suggests that we ought to build a third connotation into our use of the word 'norm'. We ought to stipulate that in order for a regularity to count as a social norm, it should not only be instantiated as a general rule, and not only seen in general as an appropriate regularity to instantiate; in addition, the fact that it is seen as appropriate – the fact that it is approved – should help to explain why it is generally instantiated.

This characterization of social norms is fairly rough, since it leaves open a range of questions. Does a regularity count as generally instantiated if it is a regularity that applies only to those holding a certain office or meeting a certain qualification? How extensive is the pattern of approval envisaged when it is said that the regularity must be seen as appropriate: appropriate morally, or at least prudentially, or at least for someone with this or that goal in mind? Moreover, must the approval be associated with the type of behavior, considered in general, or will it suffice if it just happens instance by instance that the behavior is seen as appropriate? And what, finally, is required for the pattern of approval to help to explain the pattern of behavior? Must it contribute in some measure to the production of the behavior, at least among a number of those complying? Or will it do if it is there to reinforce the behavior, should the motives that normally produce it fail for one or another reason? Will it do, in other words, if it is a virtual or standby force that is triggered to support the behavior only on a need-to-act basis, when the 'red lights' are illuminated?[2]

I am happy to leave aside most of these questions, taking the inclusive view that we should be ready to describe as a social norm any of the large range of regularities in people's behavior that meet some version of our three conditions. Thus a regularity among the members of a society will constitute a norm just in the event that:

- nearly everyone conforms;
- the behavior is nearly always thought appropriate in some way;
- and this attitude helps to explain the general conformity.

I am happy also to leave aside the further question, often raised in this context, as to whether it is essential for a social norm that the fulfillment of the three core conditions is a matter accessible to common knowledge. It will be commonly known that the conditions are fulfilled if each is aware that the conditions are fulfilled; each is aware that each is aware of this; and so on indefinitely, for any question of higher awareness that may arise. I am content that a regularity may count as a social norm even if it is not accessible to common knowledge in this sense. The upshot is a generous, perhaps deflationary sense of social norm. But it is supported by most recent authors on the subject and ought not to generate any deep controversy.[3]

[2] Pettit (1995).

[3] Hart (1961); Winch (1963); Coleman (1990); Sober and Wilson (1998); Elster (1999).

Norms in this sense often have very welcome effects. They dictate the ways in which people relate to one another in discourse, generally embracing patterns of honesty, trustworthiness and sincerity; the ways in which they otherwise seek to influence one another, eschewing resort to violence, theft, fraud, and coercion; the ways in which they commit themselves conscientiously to various collaborative causes, playing the part that is collectively required of them; and the ways in which they conduct their business and professional lives according to relevant codes of practice.

Such norms are the motors of civil society, leading people to deal well with one another, even when they are beyond the reach of the law, are unconstrained by the discipline of self-interest, and are free of the incentives provided by family and related ties. They exemplify the idea of the rule of obligation introduced by H.L.A. Hart in his classic study of 'The Concept of Law'.[4] He characterizes such rules by the fact that they generally prevail in the relevant group, they are supported by serious social pressure, they are thought useful in some way for the life of the group, and they are individually burdensome, however beneficial in group terms.

But norms in the sense defined – the sense that answers to our rough account – may also be socially neutral or even socially counterproductive, unlike Hart's rules of obligation. Norms that are socially neutral will include the sorts of norms that barely enter consciousness, such as those governing eye-contact and turn-taking in conversation, and the distance at which it is appropriate to stand in relation to an interlocutor.[5] Socially counterproductive norms come in a variety of shapes. Some will serve subcultures well, for example, while serving the society as a whole badly. Some will have a mutually destructive effect as in the norms whereby it becomes obligatory for people to exact revenge in kind for any harm done to a member of their family.[6] And some will impose fashions and fads on people who would generally prefer not to be motivated to embrace them; they represent a tyranny of majoritarian esteem.[7]

One final comment on this definition of norms, before turning to the possibility of norms sustained in error. There is no conflict in saying that a regularity that is a matter of law, being supported by legal penalty, is also a norm; it can satisfy the conditions for being a norm while satisfying also the conditions for being a law. And neither is there any conflict in saying that a regularity that is a matter of convention, say because it solves a coordination predicament to everyone's satisfaction,[8] may also be a norm; again it can consistently satisfy the conditions for belonging in each of those categories. The class of norms to which our definition directs us can overlap, and surely does overlap, with those distinct categories.

[4] Hart (1961), pp. 84–85; see also Ullmann-Margalit (1977), pp. 12–13.

[5] Goffman (1975).

[6] Elster (1990); Hardin (1995); Nisbett and Cohen (1996).

[7] McAdams (1997). See also Dharmapala and McAdams (2001) and McAdams (1995).

[8] Lewis (1969).

2 The Possibility of Error-Dependence

The definition of norms just given allows for a variety of normative regularities. Norms will vary, for example, depending on factors like the following:

- whether compliance is approved, non-compliance disapproved, or both conditions hold;
- whether the approval involved is a matter of egocentric comfort or advantage or engages a more moralized disposition;
- whether the approved compliance involves successfully achieving something or just trying to do so, as in a norm of aspiration.

I am interested here, however, in a different source of variation in norms. Under the definition offered, it is required that there is a general pattern of approval and that this helps to explain the compliance of each with the regularity in question. But there are a number of ways in which the general pattern of approval may connect with individual compliance and, correspondingly, there are different sorts of norms that the approval may support.

The standard connection, as we might describe it, involves each individual in internalizing the general pattern of approval – coming to share the attitude of approval present in the society at large – and in being led by his or her personal approval into instantiating the regularity. When such an internalized norm is in place, those who conform to it will do so out of personal attachment or conviction to the value in question and will count, by traditional criteria, as agents of virtue.

Some theorists give almost exclusive attention to internalized norms, emphasizing how far internalization supports the stability of normative behavior.[9] But even if internalization aids stability, it seems scarcely deniable that virtuous, norm-observant behavior may often be supported by something distinct. It may be reinforced – it may be protected from weakness of will and the like – by the fact that if someone does not act in accordance with a received norm, particularly a norm that is socially beneficial, then he or she is going to lose the approval of others, and perhaps attract their disapproval. Assume that for whatever reason, intrinsic or instrumental, people care about enjoying the positive approval of others and avoiding their disapproval.[10] In that case it should be clear that if failures of virtue lead to failures of behavior, then people are liable to be punished, however involuntarily, by those others who observe and understand what they have done; they will be punished, at the least, by the withdrawal of approval, or the appearance of disapproval.

Once we recognize that people may be motivated by the pursuit of others' approval, rather than by the internalization of that approval, then we are in a position to see that not all norms, however widely instantiated, may be internalized. There are two varieties of non-internalized norms that are possible, in particular.

[9] Cooter (1994; 1996).

[10] Pettit (1990); Brennan and Pettit (2004).

Given that people are motivated by the desire for the approval of others, the first possibility is that each acts out of that desire, and out of that desire only, without anyone actually having the attitude of approval in question. Everyone, in other words, is mistaken in thinking that others generally approve of a certain form of behavior – they are mistaken about the relevant value adopted by others – yet everyone displays that behavior, believing falsely that this will attract the approval of others. This possibility has been described by psychologists as a case where pluralistic ignorance supports a norm.[11]

A good example of a norm supported in pluralistic ignorance – better, perhaps, pluralistic error – is that whereby, according to one study, a group of students tended to comply with a certain regularity in the amount of alcohol consumed on a night out. The study revealed that almost all of the students disapproved of the relatively high level of drinking required under the regularity but abided by that regularity because of mistakenly thinking that others disapproved of lower levels of consumption.[12]

Should a regularity supported in such error about the values of others count as a norm, under our definition? It may seem not, since the definition presupposes that the behavior involved attracts approval, or its absence disapproval – that it actually answers to an espoused value – and all that is available in this case is the mistaken expectation of approval or disapproval. But if I believe that everyone else approves of some behavior, even a sort that I myself don't approve of, there is a sense in which I believe that that behavior is appropriate or valuable: it is appropriate-according-to-local-standards. So perhaps this regularity should count as a norm under the definition offered. Even if that is not accepted, however, it is surely reasonable to recognize how close to a norm in the strict sense this sort of regularity is and to treat it as a limit case. And if that seems too relaxed a view to take, we can go back and define a norm so that what is required is not necessarily general approval of the behavior involved but the general expectation that there will be such general approval.[13]

So much for the first sort of non-internalized norm. The second variety has not been explicitly recognized in the same way but becomes visible once we recognize that people care about enjoying the approval of others and avoiding their disapproval: for short, care about their approval. It involves error, like the first, but error about a somewhat different issue from that of whether others approve of a certain behavior. In this case, people believe that others approve of the relevant type of behavior, and they are right to do so: they get the values of others right. But here they mistakenly think that it is this personal approval that leads others, in fidelity to their values, to display that behavior. They rightly believe that the others see the behavior as appropriate or valuable, even perhaps required, but they wrongly think

[11] Miller and Prentice (1994, 1996).

[12] See Prentice and Miller (1993); Schroeder and Prentice (1998).

[13] Brennan and Pettit (2004), pp. 267–8.

that others perform the behavior out of a recognition of its merits in that regard; they mistakenly think that the others are virtuous.

This scenario of error about the virtue of others, as distinct from error about their values, is possible under the hypothesis that people care about the approval of others. For while everyone thinks that others are displaying the normative behavior out of personal virtue – as a result of their personally approving of the behavior – the fact may be that they display that behavior out of a desire to enjoy the approval of others or avoid their disapproval. The scenario is one in which everyone is 'continent' rather than virtuous, doing the right thing for reasons other than that it is right: doing the right thing for the sake of the approval of others. A norm is sustained in people's behavior but it is not internalized by anyone.

The first case of a non-internalised norm not only involved error; it involved error essentially. Did students in the drinking example know that others don't actually hold by the pattern of approval imputed by each, that knowledge might be expected to undermine the norm; and this appears to have been borne out in that actual study. Adherence to the norm in this case is, as we might say, error-dependent: more exactly, it is dependent on the near universality of the error.

Does the second sort of non-internalized norm involve error essentially, in the same way as the first? Is compliance with this norm also likely to be error-dependent? I believe that there is a possible scenario in which it would be.

Imagine that the soldiers in a military unit all display courage in action, that they all approve of courageous behavior and that those two facts are correctly registered in common belief amongst them: each believes that each believes this, and so on. Imagine, next, that they each hold by the equally common belief that the courageous behavior of others testifies to the presence of personal courage – that is, virtue – and that others act as they do out of such courage. But imagine, finally, that this belief is false, since they each act out of the belief that if they act courageously they will be taken to be courageous like the others and that they will attract approval on that account: they will avoid the stigma of being seen as cowards.

In such a situation compliance with the norm is certainly error-dependent: that is, dependent on the near universality of the error. For suppose that this error was not in place, so that it was a matter of common belief that no one is really courageous or virtuous and that they all behave courageously, if they do, out of a wish to be seen as courageous. In that case, not only will there not be a motive of courage to drive the soldiers to behave courageously, neither will the motive of seeking a reputation for courage be present; or neither at least will it be effective. For no one will be able to think that others will take them to be courageous just because they behave courageously. Behaving courageously will not be supported, then, in the motivations of the parties and, short of an awareness of this absence of motive generating some further transformation – which it well may do – courageous behavior will vanish. We can imagine an information-cascade in which the awareness that there is no reputational gain in acting courageously – in particular, no shame in failing to do so – rapidly spreads to the point where almost no one remains disposed to act that way; there is mutiny in the ranks.

One final comment. In both the cases we have considered the error committed is made by everyone. That is a particularly dramatic and clear-cut sort of situation but it is worth mentioning that error-dependence may also come in degrees. A norm may come to be maintained, not in virtue of universal error on some matter, but in virtue of a certain level of error, whether about attitudes or dispositions. In this paper I abstract from a consideration of such cases, however, as they would take me too far afield. I focus on the more clearly paradoxical case where everyone is affected by error and is led by that error into conforming with a norm. This focus is dicated primarily by a concern for keeping things simple and manageable.

3 The Plausibility of Error-Dependence

These cases of error-dependent norms may seem, on the face of it, to represent logical but very implausible scenarios. For, it may be asked, how could people be led into the egregious sort of error required for the norms to emerge or at least stabilize? Isn't it willful and *ad hoc* to postulate such a rank degree of proneness to error about the values or virtues of others? I shall try to argue that it is not.

Social psychologists have documented a consistent tendency in people's attribution of motives to others that they describe as the fundamental attribution error; I prefer to think of it as a bias, since it may not invariably lead to mistakes. E.E. Jones presents it in the following terms. 'I have a candidate for the most robust and repeatable finding in social psychology: the tendency to see behavior as caused by a stable personal disposition of the actor when it can be just as easily explained as a natural response to more than adequate situational pressure'.[14]

The fundamental attribution bias consists in a preference for explanations of what people do, and of what people say, that emphasizes the contribution of character over context. Let someone do or say something and there will be many possible explanations. Some will suggest that people adapt in very finely tuned ways to differences of circumstances, so that there are no easily predictable patterns in evidence. Others will depict people as possessed of dispositions that are stable across different contexts and that dictate predictable patterns of response: patterns that materialize reliably across quite different sorts of circumstances. The fundamental attribution bias consists in a preference for this second sort of explanation, in which character receives relatively more weight, context relatively less.

The finding about the fundamental attribution bias – and I shall assume that it is a well-documented finding – bears in an obvious way on issues of the kind that arise with non-internalized norms.

Suppose that your drinking companions routinely drink a certain amount and routinely acquiesce in a general acceptance of that level of drinking, perhaps even offering general applause for going that far: we may well expect such applause, given that under the story presented it too can be expected – mistakenly – to earn

[14] Jones (1990), p. 138.

approval. What more natural explanation of your companions' behavior, then, than that they approve of that level of drinking – in particular, perhaps disapprove of a lower level – and that this approval plays a role in explaining their behavior? To explain what they do in those terms is to project on them a stable complex of attitude and motive. The disposition imputed suggests that regardless of who they are drinking with, they should each be expected to maintain the same level of consumption; it is to downplay the role that circumstances might be expected to play.

Contrast this sort of explanation, however, with that which appeals to the desire of each to be well regarded by others. Under that account, each has to explain the behavior of the others as the product, not of a simple attitude-motive complex, but in terms that give a much larger role to context. Each will take others to be stable of disposition at the high level of seeking the approval of others. But the disposition imputed will be expressed in action in very different ways, depending on differences of circumstances. Let the context be one where others are expected to approve of a high level of drinking, and they will tend to drink to that level. Let it be one where others are expected to disapprove of such drinking, and they will tend to drink less.

The fundamental attribution bias is bound to lead people to explain the presentation of others in the simpler, more context-invariable way. And it is precisely that sort of explanation that each is expected to give of the behavior of others in the story about the fixing of the drinking norm. True, the explanation given turns out to be a mistake. But it is not an implausible explanation to posit. It is the sort of account that is bound to appeal to anyone who is susceptible – as we are all said to be susceptible – to the fundamental attribution bias.

The comments made on this case carry over to the second case of non-internalized norms. Imagine that you are a soldier among soldiers, involved in a series of dangerous military engagements. There is no doubt in your mind but that you and your fellows all approve of courage. Nor is there any doubt about this being a matter of shared awareness. After all, you all talk about courage in the most positive terms, and listen to regular harangues on the topic. So now you find that the others in the ranks with you do indeed behave very courageously. What more natural explanation than that this is due to the fact that they put their lives where their words are: they march to the drum of their values?

This explanation is straightforwardly in line with the fundamental attribution bias, imputing to others a stable disposition – again, a simple attitude-motive complex – that can be expected to produce similar behavior across many fine differences of context. Contrast with that explanation the account that you would have to give of others' behavior, were you to grasp the facts of the case, as my hypothesis presents them. As in the other case, you would posit a stable, high-level disposition – the concern with approval – that is liable to produce quite different behaviors in different contexts. You would downgrade character and upgrade context in a way that runs deeply against the grain of the fundamental attribution bias.

I conclude that in these respects the stories told in illustrating the possibility of error-dependent norms are not at all implausible. The alleged mistakes about the values and virtues of others are precisely the sorts of error that we should expect them to be liable to make, according to received psychological analysis.

But there is a second way in which the stories told may be thought to be implausible. They suppose that people can be moved by one incentive but can expect at the same time that others will explain their behavior by reference to another. The motive that actually operates, according to those stories, involves the desire to enjoy the approval of others and escape their disapproval. But the motive that agents expect to be ascribed involves a much lower-level, context-insensitive disposition. Is it reasonable to suppose that people could expect others to make that sort of mistake?

I believe it is, again because of the presence of the fundamental attribution bias. The motive people expect to be attributed is the disposition to act in a way that answers to how they are (really or apparently) disposed to approve and disapprove of behavior: in the one case the disposition to drink to the level they make a show of endorsing; in the other the disposition to act courageously after a pattern they clearly admire. As there is no surprise about people's explaining the behavior of others in the erroneous manner described, then, so there should be no surprise about their expecting others to make the same mistake in relation to them. If the fundamental attribution bias is really a feature of human psychology, then it is likely to be a salient feature and one that people are likely to expect others to display.

One final question. On the account given, people explain others' behavior by ascribing a stable, dispositional motive, and they expect others in turn to explain their behavior by reference to a similar motive. But they themselves, according to our account, act on the basis of a pattern of motivation that gives the lie to the notion that low-level, dispositional motives are the more or less ubiquitous sources of action. And they may more or less consciously act this basis. Does this make the account inconsistent? Or does it at least mean that those to whom the account is applied are inconsistent?

No, it does not. I might consistently think that I am motivated differently from others, that others don't see this, and that actually I am exceptional; there is independent evidence, indeed, that this is how we each tend to think of ourselves, avoiding the attribution bias in our own case.[15] Unlike others, so I may feel, I lack the required attitude of approval, or I lack the disposition to act as I approve. But I do of course see the importance of securing approval and so fall back myself on this other motive. On the account offered, we might suggest that people generally think they are different from others in these respects. And that would not be a particularly difficult hypothesis to embrace, given the frequent asymmetry in how people view themselves and others.

There is a second fairly plausible hypothesis that would also explain how I could act on one motive but expect to be ascribed another. This is that I am not always aware of how my own motives work; it is not always the case that I more or less consciously act out of a desire for esteem or a fear of disesteem. While being sensitive primarily to the forces of approval and disapproval, I may often imagine, whether out of self-ignorance or self-deception, that I am not like this: that I am, in fact, exactly like others in being moved by the same values or the same virtues.

[15] Jones and Nisbett (1971).

To sum up, then, there is no deep implausibility in holding by the assumptions that would predict and explain the appearance of error-dependent norms. The fundamental attribution bias, powerfully supported as it allegedly is, should lead us to expect exactly the pattern of motivational attribution that would account for the presence of such norms.

4 Varieties of Error-Dependent Norms

There are two broad types of error-dependent norms identified in the story told so far, one sort exemplified by the drinking case, the other by the case of military courage. The first I describe as value-mistaken norms, the second as virtue-mistaken norms.

Value-mistaken norms will be liable to emerge only so far as two conditions are fulfilled. The first is that those among whom the norm emerges care greatly for one another's approval; in particular, care enough to be ready to act against their own values for the sake of the approval of others. And the second is that conditions are such as to allow the parties to be mistaken about one another's values, and so about what they are each likely to approve or disapprove.

The first condition suggests that there will be considerable peer pressure associated with membership of the group in question. And the second implies that members of that group do not come together out of a search on the part of each for those of a common mind: a search for people who share the same values and the same attitudes of approval and disapproval; the group forms on some other basis. These conditions are likely to be fulfilled quite commonly, so that we should not be surprised if there are a variety of value-mistaken norms in place in any society.

Two salient possibilities are worth mentioning. One involves what we might think of norms of corruption: that is, corruption from the point of view of the wider society. Take a group of police officers who have a defensive attitude towards outsiders, expecting them to be somewhat hostile. Membership and acceptance in the group will tend to matter greatly to officers, particularly so far as they expect outsiders not to accept them fully. Given that they do not invariably join the force out of any particular values, there will be room in such a group for members to be unsure about what exactly others care about. In such a scenario it is very possible that a number of value-mistaken norms may emerge or stabilize.

Imagine that someone in the group is seen by others as breaching police rules – that is, the rules imposed by higher authority – in some perhaps not very significant manner. Will any of those others report the breach? Very possibly not, since they are each liable to be afraid of alienating themselves from the offender and his or her friends; and this, even if they each actually disapprove quite strongly of the breach. But suppose now that this pattern of not welching on fellow-officers begins to get established as a routine of behavior. As it does, this will suggest to each that others disapprove of welching and approve of turning a blind eye to an individual officer's offences, at least within certain limits. And so we can imagine that a norm of not welching may emerge, with almost everyone conforming, with almost everyone expecting almost everyone to approve of conformity, and with this expectation helping to keep the norm in place. But for all that our story entails, the norm may be value-mistaken in character and so not internalized by the parties. The members of the force may each disapprove of not welching – they may embrace the value of whistle-blowing – and may each keep to the pattern of not welching only out of an erroneous expectation as to how others would respond to whistle-blowing.

If there is a norm of not welching in place among the members of such a police force, of course, then there will be room for the emergence of other value-mistaken norms as well. Suppose that while members of the group think that others disapprove of whistle-blowing, they believe that there are certain limits to what will be tolerated: there are forms of breach such that no one would disapprove of their being publicized. And suppose now that a certain form of breach that goes beyond the ascribed limits begins to emerge in the group – say, one of taking bribes – yet is not exposed by fellow officers. It may be that each is wrong in thinking that this is beyond the limits of tolerance; it may be that the behavior is reported by no one, out of fear of incurring disapproval as a whistle-blower. But so far as there is a general mistake made about the limits of tolerance – surely, a real possibility – each may then be led to think, again mistakenly, that actually the behavior in question is not disapproved of by others, perhaps even that it attracts a degree of approval. The error about attitudes to whistle-blowing can generate other errors too. Value-mistaken norms may multiply and propagate.

The group of police officers imagined exemplifies a range of real-world possibilities. Instead of police officers we might have imagined the lower-level workers in any enterprise, the students in a school, the inmates in a prison, or the members of any professional association. In every such case there will be motives in place that allow for a value-mistaken norm of not whistle-blowing to emerge and, as a result, for the appearance of other value-mistaken norms. These other norms may lead to a tolerance of shirking or bullying or closing ranks against outsiders, even when everyone in the relevant group disapproves of those behaviors.

I do not say, of course, that whenever there is a norm against whistle-blowing, or whenever any of the associated norms prevails, that is because people are mistaken about one another's values. But I do say that even if individuals hold values that would support whistle-blowing and would outlaw a variety of other breaches, those values need not give rise to corresponding norms.

The value-mistaken norms illustrated can be described as norms of corruption, since they support behaviors that fall away from standards that others expect members of the group to honor. But it is worth mentioning that they do not exhaust the possibilities for value-mistaken norms. A second, more or less salient possibility is that norms of correctness, as I will call them, may emerge on the basis of mistaken values. I take the word 'correctness' from the way in which we speak of political correctness, though I believe that norms of correctness may run far beyond the limits of any explicitly political context.

With norms of corruption people are misled by the action or inaction of others into ascribing values that none of them actually endorses. With norms of correctness, they will tend to be misled by the speech or silence of others into making similar mistakes. Suppose that a group of a political or religious or cultural character is established and, as in the earlier sort of case, that it commands great allegiance among members; they each care greatly about belonging. And now imagine how people are likely to respond to the words of some authorities or would-be authorities within the group, when they declare what the attitudes and values of the group are and what they require. Such a declaration will carry influence, so far as no one opposes it. Yet people may each fail to oppose it, not because of sharing the values declared, but because of assuming that others do share them and because of fearing the ostracism that would go with resisting the common line. No one may oppose the declaration, in short, because no one may dare oppose it; no one may think that the risk of being the only dissident voice is worth taking.

Nor is this all. It is also possible in such a case that the person or persons who assume the role of authority and declare the values of the group are not themselves sincere. Just as those who fail to oppose them may not share in the declared values, so the authorities themselves may fail to share in them. The people involved may mistakenly think that others do share the values and may put themselves forward as spokespersons for those values, not out of an attachment to the ideals, but out of a wish to ingratiate themselves with other members of the group.

This possibility is all too easy to envisage. Any number of political and religious and social movements are liable to generate the sort of pressure under which words can engender norms that no one dares to dishonor, even while they are each opposed to the norms in their hearts. Insincerely endorsed words may be cheap for those who utter them; indeed, they may promise a positive reward in the approval that they are expected to earn. And insincerely accepted or unopposed words may be powerful; they may give rise to a cascade of behaviors that no one approves but that everyone displays.

So much for value-mistaken norms, whether of corruption or correctness. The other error-dependent norms that we identified are virtue-mistaken rather than value-mistaken. They arise, not on the basis of an error about what others approve or disapprove, but on the basis of an error about why others display the behavior of which they approve and avoid the behavior of which they disapprove. The error here consists in thinking that others are virtuous – they are disposed to act according to their values – and that it is their virtue rather than any ulterior motive that explains why their actions conform to what they say and think.

Virtue-mistaken norms presuppose that membership and acceptance matter to people just as much as they do in the other case, so that each is loathe to lose the approval of others or earn their disapproval. But it is a matter of common awareness in this case that such and such patterns of behavior are approved of or disapproved of – the values of people in the group are manifest. Here error can kick in, then, not in the perception of how others approve or disapprove, but only in the perception of their motives. Why, however, might it kick in? Why might it be the case that although everyone thinks that others are possessed of the virtue of living up to their values, no one actually has that sort of virtue?

The obvious answer is that while such virtue is a natural disposition to ascribe in explanation of value-conforming behavior – this, in view of the fundamental attribution bias – the value in question is very demanding and is unlikely to be capable of motivating the ordinary run of people, all on its own. In order to identify examples of virtue-mistaken norms, then, we need only reflect on cases where the values that people proclaim – and proclaim sincerely – are really very demanding and are unlikely to have such an imaginative and emotional appeal that they will routinely gird agents against temptation. People will see others generally conforming but will have to struggle against temptation to conform themselves. And when they succeed, if they succeed, that will not be because they are as virtuous as they think others are. It will be because the cost of failure, as they see it, would be too grave to bear: it would consist in being seen, to their shame, as more or less isolated defectors.

The case of military courage exemplifies this structure faithfully. It goes against the grain of human nature, by the testimony of history and introspection, to expose oneself to grievous physical danger; or to do this, at any rate, when there is no immediate goal like that of aiding a comrade in trouble. But in the military context everyone will clearly regard action in face of such danger as a supreme value and everyone will see others as having enough virtue – enough courage – to be able to honor that value in practice. In such a context, the prospect of being the only one not to act virtuously – the only one not to act as if they had the virtue – will promise ignominy and stigma. It should be no surprise if, under the force of that motive, everyone does indeed act virtuously. No one has the virtue of courage: no one has an attachment to the value that is strong enough to see them through adversity. But, this being the only way to avoid the prospect of a crushing burden of shame, everyone simulates the presence of such virtue and seeks to hide what they wrongly see as their untypical, perhaps unique, cowardice.

The structure that courage exemplifies, according to this story, is liable to be replicated across many different contexts. Take any of the values that are universally acknowledged, difficult in practice to honor, yet fairly generally conformed to. In any such case, there will be a possibility that the norm is maintained on a virtue-mistaken basis. There is going to be a question, then, as to how far various norms may not be due to mistakes on people's parts about one another's virtue; in particular, mistakes sourced in a credulous acceptance that others are made of better stuff than they are themselves.

We have focused in this essay, for simplicity, on cases where error is wholesale rather than coming in degree. There may not be many norms where it is plausible to think that they are virtue-mistaken in that wholesale way but there are likely to be many where such error plays at least a partial role. Think of values that are commonly proclaimed, whether in society as a whole or in certain groups, such as punctiliousness in making one's tax return, steadfastness in upholding a religious faith, or conscientiousness in preparing oneself for committee deliberations. All of these values make demands that may prove hard to sustain as one is assailed by the temptation to fudge some financial details, to raise doubts about a church's teaching, or to cut corners in reading background material. People may well overcome such temptations by dint of attachment to the relevant value. But there is always a possibility that the real force that enables them to achieve this success is not their natural virtue but the perhaps credulous belief that others do have such virtue combined with the desire to avoid the shame of appearing to be more or less isolated deviants.

The examples of value-mistaken and virtue-mistaken norms that we have canvassed in this last section should be sufficient, I hope, to show that error-dependent norms are not only a possibility; it is very likely that some such norms obtain in any society or any grouping. That people conform to norms, then, does not mean that those norms are internalized amongst them: that they internalize the values and the corresponding virtuous dispositions. Many norms may prevail – whether for good or for ill – as a result of the simulated adherence to certain values or the simulated display of certain virtues. The social patterns that obtain among people in aggregate may give only very misleading cues as to the patterns that obtain in their individual souls.

5 Towards Institutional Design

If there is likely to be such a cleavage between patterns in social life and the patterns in people's souls, of course, then that raises a serious question as to how far we should deplore this phenomenon, or how far we should try to turn it to good. The moralist's response will be to say that we should deplore it. But the institutional designer's will be that, on the contrary, we should try to build on it. Recognizing that the task in institutional design is to construct something straight out of what Kant called the crooked timber of humanity, we should try to see how far mistakes about value and virtue can be exploited for social good.

I favor the response of the institutional designer. It may seem that the response involves condoning and encouraging hypocrisy. But the hypocrisy involved is special. Hypocrisy in the normal sense means dissimulation: pretending to have done something approved of, or to have avoided something disapproved of, when this is not actually the case. Hypocrisy of the sort displayed in upholding error-dependent norms is quite different. It means doing something good out of a love of approval, or out of a fear of disapproval, while pretending to do it out of a love of the value relevant to the approval or disapproval. This is not the hypocrisy of dissimulation but, in the old phrase, the hypocrisy of simulation: the hypocrisy involved in simulating the behavior of those who internalize a norm, without having internalized it oneself. Simulation is a very benign sort of hypocrisy, since it leads to the same behavior as internalization of the norm; the shortfall associated with it is not behavioral but only motivational.

Assuming that we should not baulk at the prospect of condoning such simulation, what steps in institutional design would the observations of this paper support? The paper supposes that people care about the esteem of others, and shrink from their disesteem; and that mutual exposure is sufficiently assured to engage this concern and motivate the simulation of internalized, normative behavior. But it directs our attention to two phenomena in particular that may be exploited in mobilizing the forces of esteem and using them to generate or reinforce socially desirable patterns of behavior; I shall assume for simplicity that there is no issue about what patterns count as desirable, what not. The first is the tendency of people to be guided by a desire for correctness, in particular correctness according to what are taken as the expectations of a group. And the second is people's credulity in being prepared to accept that others are virtuously attached to a value that they themselves find it very hard to internalize in their motivations.

We can easily imagine situations in which the culture of correctness gives life – and so, can be exploited to give life – to some socially attractive norm. Without really recognizing or internalizing the value of not smoking in the company of others, or putting one's recycling in a separate bin from one's trash, or using language that respects proprieties of gender or race, it seems entirely plausible that people should be moved to act in a way that supports such patterns. But there is no block to this being true of everyone. And so we can equally imagine that while each acts in a way that will attract the approval of those who recognize the appropriate value, actually there may be few if any members of the society who do recognize it.

The norms we imagine emerging under a culture of correctness, we can equally easily imagine emerging as a result of people's credulity in thinking that others are more virtuous than they are themselves. Consider the old emphasis on the importance of giving a good example to others, in particular giving a good example to those who might be expected to be less attached to a relevant norm. This underscores the common belief that if people have faith in the virtue of their fellows – including the virtue of their betters – then this sharpens the incentives they have to simulate that virtue, thereby winning esteem or avoiding disesteem. The belief is also underscored by the common fear of revealing to the members of a community just how many of them fail to live up to certain standards. It is a rare society in which government will be happy to reveal the fact that fewer people comply with the tax laws than is generally believed. It is a rare church in which the authorities will concede that fewer members conform to certain norms of faith or morals than might be generally expected. This being so, there can be little doubt about the possibility of exploiting people's credulity in order to muscle certain norms into existence.

But could we really hope to establish a norm by putting it about that correctness requires such and such behavior; or by making it seem that many people act out of a love of the value involved in the norm? Wouldn't any norm established in that way be very unstable, being susceptible to collapse at the point where the truth comes out: the Emperor has no clothes? I don't think so. Norms that begin in a culture of correctness or a habit of credulity can gain a more solid hold if the value involved – unlike perhaps the value of becoming intoxicated with others – is one that can really engage human affections. The point is particularly persuasive if we consider a case, not where absolutely no one holds the value at the start, but where a portion of the society do. Correctness and credulity may cause the class of those who hold by the value to increase, as more and more of those who pay it lip service only – behavioral service only – come to register and feel its affective grip.

This claim should not be surprising. There is a long and plausible tradition of thinking that internalized adherence to norms can be supported and stabilized by collateral incentives and my claim about the power of a culture of correctness, or a habit of credulity about the virtue of others, fits easily within this. The tradition may have its origin in the Aristotelian idea that continence – compliance with the right for reasons independent of its rightness – can support and even give rise to virtue proper: to a compliance with right behavior that is a matter of second nature and bespeaks an attachment to the right.[16] Doing the right thing for the wrong reasons may be a vice: to quote from T.S. Eliot's 'Murder in the Cathedral', it may be 'the greatest treason/To do the right deed for the wrong reason'. But if it is a vice, then it is a saving vice; it represents a form of motivation and behavior that is friendly to virtue, not inimical.[17]

We should not rail against the resort to institutional design, then, not even against the resort to mobilizing correctness and credulity for design purposes. It is now widely recognized that people become responsible citizens through being 'responsibilized'.[18] If they are held responsible for certain failures on a basis of more or less strict liability, and they know that this is so, then they are likely to become sufficiently attentive to ensure that any failure which occurs will actually be a matter of fault; they will be blameworthy in the ordinary sense, not in the sense of being just strictly liable.[19] People may be made fit to be held responsible by being treated as if they were fit to be held responsible.

The reason we should not rail against institutional design of the sort envisaged here is that what is true of responsibility may also be true of normative sensitivity. As we may make people responsible by treating them as if they were responsible, so we may make people normatively sensitive by getting them to behave, out of a love of esteem, as if they were sensitive. As we may responsibilize people, so we may equally hope to be able to 'sensitize' them: that is, to sensitize them to the claims of suitable values.

This may sound like the expression of a government-house mentality, in which we the institutional designers think of ourselves as different from those on whom impose our designs. But that need not be so. We should each be aware of the complexity and fragility of our own motivations; none of us can be complacent, for example, about the possibility of remaining virtuous while wearing the ring of Gyges. And being aware of our own nature in that way, we should embrace the prospect of being scaffolded in our fragile commitment to relevant values. We should embrace this prospect even as we recognize that that scaffold may be the product of human design, and perhaps the product of our own designing initiatives.

[16] Aristotle (1976).

[17] Lovejoy (1961); Brennan and Pettit (2004).

[18] Garland (2001).

[19] Pettit (2001), Chap. 1.

References

Aristotle (1976): *The Nicomachean Ethics*, Harmondsworth: Penguin.

Brennan, G. and Pettit, P. (1993): "Hands Invisible and Intangible", in: *Synthese* 94: pp. 191–225.

Brennan, G. and Pettit, P. (2000): "The Hidden Economy of Esteem", in: *Economics and Philosophy* 16: pp. 77–98.

Brennan, G. and Pettit, P. (2004): *The Economy of Esteem: An Essay on Civil and Political Society*, Oxford: Oxford University Press.

Coleman, J. (1990): *Foundations of Social Theory*, Cambridge (MA): Harvard University Press.

Cooter, R. D. (1994): "Structural Adjudication and the new Law Merchant: A Model of Decentralized Law", in: *International Journal of Law and Economics* 14: pp. 215–231.

Cooter, R. D. (1996): "Decentralized Law for a Complex Economy: The Structural Approach to Adjudicating the New Law Merchant", in: *University of Pennsylvania Law Review* 144: pp. 1643–1696.

Dharmapala, D. and McAdams, R. H. (2001): "Words that Kill: An Economic Perspective on Hate Speech and Hate Crimes", in: *University of Illisnouis Law and Economics Research Papers*. Champaign: Urbana.

Elster, J. (1990): "Norms of Revenge", in: *Ethics* 100: pp. 862–85.

Elster, J. (1999): *Alchemies of the Mind: Rationality and the Emotions*, Cambridge: Cambridge University Press.

Garland, D. (2001): *The Culture of Control: Crime and Social Order in Contemporary Society*, Chicago: University of Chicago Press.

Goffman, E. (1975): *Frame Analysis: An Essay on the Organization of Experience*, Harmondsworth: Penguin Books Ltd.

Hardin, R. (1995): *One for All: The Logic of Group Conflict*, Princeton (N.J.): Princeton University Press.

Hart, H. L. A. (1961): *The Concept of Law*, Oxford: Oxford University Press.

Jones, E. E. (1990): *Interpersonal Perception*, New York: Freeman.

Jones, E. E. and Nisbett, R. E. (1971): *The Actor and the Observer: Divergent Perceptions of the Causes of Behavior*, New York: General Learning Press.

Lewis, D. (1969): *Convention*, Cambridge (MA): Harvard University Press.

Lovejoy, A. O. (1961): *Reflections on Human Nature*, Baltimore: Johns Hopkins Press.

McAdams, R. H. (1995): "Cooperation and Conflict: The Economics of Group Status Production and Race Discrimination", in: *Harvard Law Review* 108: pp. 1003–1084.

McAdams, R. H. (1997): "The Origin, Development and Regulation of Norms", in: *Michigan Law Review* 96: pp. 338–433.

Miller, D. T. and Prentice, D. A. (1994): "Collective Errors and Errors about the Collective", in: *Personality and Social Psychology Bulletin* 20: pp. 541–550.

Miller, D. T. and Prentice, D. A. (1996): "The Construction of Social Norms and Standards", in: *Social Psychology: Handbook of Basic Principles*, E. T. Higgins and A. W. Kruglanski (eds.), New York: Guilford Press: pp. 799–829.

Nisbett, R. E. and Cohen, D. (1996): *Culture of Honor: The Psychology of Violence in the South*, Boulder: Westview Press.

Pettit, P. (1990): "Virtus Normativa: A Rational Choice Perspective", in: *Ethics* 100: pp. 725–755; reprinted in: Pettit, P. (2002): *Rules, Reasons, and Norms*, Oxford: Oxford University Press.

Pettit, P. (1995): "The Virtual Reality of Homo Economicus", in: *Monist* 78: pp. 308–329; Expanded version in Maki, U. (ed.), (2000): *The World of Economics*, Cambridge: Cambridge University Press; reprinted in Pettit, P. (2002): *Rules, Reasons, and Norms*, Oxford: Oxford University Press.

Pettit, P. (2001): *A Theory of Freedom: From the Psychology to the Politics of Agency*, Cambridge and New York: Polity and Oxford University Press.

Prentice, D. A. and Miller, D. T. (1993): "Pluralistic Ignorance and Alcohol Use on Campus", in: *Journal of Personality and Social Psychology* 64: pp. 243–256.

Schroeder, C. M. and Prentice, D. A. (1998): "Exposing Pluralistic Ignorance to Reduce Alcohol Use Among College Students", in: *Journal of Applied Social Psychology* 28: pp. 2150–2180.

Sober, E. and Wilson, D. S. (1998): *Unto Others: The Evolution and Psychology of Unselfish Behavior*, Cambridge (MA): Harvard University Press.

Ullmann-Margalit, E.·(1977): *The Emergence of Norms*, Oxford: Oxford University Press.

Winch, P. (1963): *The Idea of a Social Science and its Relation to Philosophy*, London: Routledge.