Bibliography

Akerlof, G. (1984), A Theory of Social Custom, of which Unemployment May Be One Consequence, in: *An Economic Theorist's Book of Tales*, Cambridge, 69–100

Bedau, H. A. (1961), On Civil Disobedience, in: *Journal of Philosophy 58*, 653–665

— (1969) (ed.), *Civil Disobedience*, New York

Brams, S. J. (1978), *The Presidential Election Game*, New Haven

Cohen, C. (1971), *Civil Disobedience: Conscience, Tactics and the Law*, New York

Dworkin, R. M. (1967), The Model of Rules, in: *University of Chicago Law Review 35*, 14–46

Fried, Ch. (1981), *Contract as Promise*, Cambridge/MA

Frohlick, N./J. A. Oppenheimer/O. R. Young (1971), *Political Leadership and Collective Goods*, Princeton

Goodin, R. E. (1990), International Ethics and the Environmental Crisis, in: *Ethics & International Affairs 4*, 91–105

— (2005), Toward an International Rule of Law: Distinguishing International Law-breakers from Would-be Law-makers, in: *The Journal of Ethics 9*, 225–46

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

Kuran, T. (1998), Ethnic Norms and Their Transformation through Reputational Cascades, in: *Journal of Legal Studies 27(2)*, 623–59

Mills, C. W. (1940), Situated Actions and Vocabularies of Motive, in: *American Sociological Review 5*, 904–13

Posner, R. A. (1972), *Economic Analysis of Law*, Boston

Simpson, A. W. B. (1973), The Common Law and Legal Theory, in: A. W. B Simpson (ed.), *Oxford Essays in Jurisprudence*, 2nd series, Oxford, 77–99

Tilly, Ch. (2006), *Why?*, Princeton

Weber, M. (1947), *Theory of Social & Economic Organization*. Trans. A. M. Henderson and T. Parsons, New York

Holly Lawford-Smith

The Importance of Being Earnest, and the Difficulty of Faking It

A Comment on Robert Goodin

1. Introduction

Goodin's *Norms Honoured in the Breach* is the companion to his earlier paper *Toward an International Rule of Law: Distinguishing International Law-Breakers from Would-Be Law Makers* (2005). In the domestic context there are set procedures for legal reform, so the straightforward way to distinguish law breakers from would-be law-makers (i.e. those seeking legal reform) would be to look at whether they break the existing laws, or instead take the appropriate steps to change the existing laws. But international law is an interesting case, because there is little by way of set procedure for reform—and in fact little by way of genuine law, if you think that enforceability is necessary to render law 'genuine'. One way to push for legal reform in a system with no set procedures for doing so is to *break* the law. But that poses an interesting question: how can we tell when a state is simply breaking international law, and how can we tell when it is pushing for international law reform? In his (2005) paper Goodin considers several answers to that question, arguing in general that would-be lawmakers will break the law publicly, accept the accompanying sanctions, and accept the legitimacy of other states acting in the same way for the same reasons. Presumably, would-be law-making is normatively acceptable, under the same conditions that conscientious objection in the domestic case is normatively acceptable.

In his article in this volume, Goodin exploits his earlier work, extending the conditions that distinguish a would-be lawmaker to cover domestic norm breakers. The paper might be read uncharitably as a handy 'how to' guide for the would-be social cheater. Goodin offers his readers some useful tips: don't break norms too often, or your excuses will start to seem disingenuous; make sure to sincerely protest your special reasons for breaking the norm, and your otherwise firm commitment to it; and it wouldn't hurt to throw in a little hand-wringing for effect. In short, the lesson is that the very behaviour that makes both international and domestic law-breaking normatively acceptable can be emulated to allow successful social defection. Very well. Goodin has adopted the perspective of the social agent concerned to cheat and get away with it; here I shall take the perspective of the society concerned to identify the threat of such ingenious defectors. The discussion shall proceed covering two central questions. Firstly, I'll ask whether there's actually anything wrong with the odd well-disguised defection. If there's not, maybe there's nothing that society needs to protect itself from. Secondly, I'll discuss the likelihood of society actually being confronted by Goodin-style cheaters. I will draw upon arguments of the evolu-

tionary biologists to suggest firstly that sincerity of the kind Goodin proposes is hard to fake, and secondly that norm-breaking, whether accompanied by justification or not, might still signal unreliability as a cooperative partner.

2. What's Wrong with a Little Defection Every Now and Then?

Whenever there is large-scale cooperation without the presence of a central authority, norms are usually wheeled in to explain it. But what is a norm? To some extent any definition depends on a theorist's purposes. Sociologists might prefer to define norms in terms of expectations, or values, while philosophers might prefer a behavioural definition. Robert Axelrod, in one of the early studies using game theory to analyze cooperation, preferred a behavioural definition on the grounds that it allowed us to infer the presence of a norm from behaviour, and it permitted our being able to say that norms were present to greater or lesser degrees (Axelrod 1984; 1986). Philip Pettit more recently defines a norm in similarly behavioural terms as "a regularity that actually prevails among members of a group" (Pettit 2008). He qualifies that by adding that the behavioural regularity must attract a high approval rating, where the high approval rating should help to explain why the norm is generally instantiated. That a regularity satisfies those three conditions should in general be a matter of common awareness.

Consider two quite different cases involving norms. In the first case, a homosexual couple choose to challenge a firmly-held social norm against public displays of homosexual affection by walking through their town centre holding hands. In the second case, a student chooses to fly from Berlin to London rather than taking the train, despite a strong social norm in favour of environmentally friendly travel. What should be immediately obvious is the different normative status of each of these acts. The homosexual couple breaks the norm against public display of homosexual affection in the hope of changing it. Most people would agree that such a norm is oppressive and lacks reasonable justification, and would therefore welcome both the challenge to it and the possibility that the challenge trigger a 'normative cascade' whereupon more and more people defect from the norm in question. In general, we think that the breaking of discriminatory or otherwise pernicious social norms is a good thing. The second case is quite different. The norm is in place for a reason, namely that countries *must* lower their carbon emissions. Moreover, the student doesn't break the norm because she particularly wants to change it. We can suppose that she doesn't care much either way, or that she prefers other people to conform to the norm so long as she can be the exception. Whatever it is, we think there's something wrong with the fact that she breaks the norm, not least because of the risk, which Goodin discusses extensively, of her defection triggering similar behaviour in others. In general, we think the breaking of fair or otherwise desirable social norms is a bad thing.

Goodin's focus is squarely upon cases of this second kind. He wants to investigate the possibility of a person breaking a norm while avoiding the risks of triggering defection in others. In game-theoretic terms, he takes the perspective of an economically self-interested player in a cooperation game. The dominant strategy in many such games (or at least the dominant strategy for someone unlikely to have future interactions with his fellow players and lacking preferences for anything other than the accumulation of economic goods) is to cheat. That strategy obtains the maximum possible payoff. But in real world terms, the defection can't be admitted to be such, because that would be to invite retaliation, or at least imitation. That would likely put a fragile arrangement such as conditional cooperation to produce some public good (or worse, and more topically, to not lose some fundamental public good which is under threat) at risk of systemic collapse. Hence the title of this section, 'what's wrong with a little defection now and then?' Obviously it's at least sometimes, if not always, permissible to break pernicious social norms; but it's not obvious that it's permissible to break the good ones. 'Defection' is a term usually reserved for the latter, i.e. social transgressions generally contrary to public or collective welfare.

Whether you think there's anything wrong with a little defection every now and then will depend largely on background views about norms, duties, and obligations. A preliminary question to ask involves the connection between social norms and moral duties. It can't be that it is morally obligatory to obey all social norms, because we've already considered cases of pernicious social norms where we think the right thing to do is to break them, in the hope of changing them. But it might be that certain social norms are grounded in the right reasons, like recently developing norms toward more environmentally-friendly behaviour. Assuming that it is morally obligatory to obey some subset of social norms, e.g. the good ones, it still doesn't follow that *everyone* has to obey them. This is where things will depend on the controversial nature of rules and obligations. If you think obligations are categorical and exceptionless then you'll think the person who breaks a positive social norm has morally transgressed. If you think rules and obligations are hypothetical and conditional, then you might not think there is moral transgression at stake in a defection. And regardless of whether you think there's a moral debt being incurred, there's a practical question of harm done.

That is to say, if we're strict consequentialists, then the important questions are about outcomes. Obviously if a person defects on a positive social norm and triggers a normative cascade such that the norm collapses, they've done potentially irreparable damage. But Goodin's point is that the would-be defector can emulate the strategies of the conscientious objector and the international law re-maker in order to avoid precisely that outcome. And if he does that, then the only damage done is the moral damage of his own transgression, which might look fairly minor compared with the alternative just considered.

The problem is that there will be cases where non-compliance with a norm creates a remainder, and cases where it does not. In cases where, say, there is a norm of praying at morning meals, the fact that one person alone stops doing this has no effect on anyone else. That kind of case is a perfect candidate for the kind of response that says a little defection every now and again doesn't matter. The only

harm accrued is to the person who broke the norm. But more often we will be dealing with cases where defection means that a person fails to do their fair share, which will entail that the rest of a community of cooperators have to absorb the cost of defection. If one person always buys the tea and coffee for the tea room, and another person never puts the agreed donation in the money jar, the person doing the buying is forced to absorb the defector's costs. If everyone in a flat agrees to go without heaters to drive the cost of the electricity bill down, and one person surreptitiously defects, the others will have to pay costs for which they receive no benefit. And so on. Where duties are divisible among persons and there is defection, the defector's duties will accrue to others. That means others must do more than their fair share, which affects their welfare compared to those who have to do a normal share, and especially compared to the defector who took no share at all. We can get around at least a subset of these cases by doing theory that assumes imperfect compliance (which anyone designing policy in the real world does, although some would argue that perfect compliance is just a matter of arranging the costs and incentives in the right way). Then the important thing is to figure out how much defection a policy can withstand, and to find ways to keep cooperation above that level. The only risk of that kind of set-up is that if people know that not everyone has to cooperate, there'll be a kind of 'race to the bottom' in order to be the one to get the fruits of cooperation without paying the price.

One final thought on this point is that rather than taking Goodin's discussion to be a 'how to' guide for the would-be defector, a more charitable interpretation might be that he hopes to persuade the would-be defector to minimize damage to collective endeavors. He tells the would-be cheater that there's more at stake than her own selfish preferences; he asks her to consider the *effects* which her cheating might have, the greatest of which is the potential normative cascade. Or indeed he tells the person who genuinely *is* an exception to a norm and is excused in violating it how to avoid being misinterpreted (this reading is consistent with his talk of 'conscientious' contrasted with 'callous' rule-breaking). This cascade was a good thing in the case of norms in need of reform, but stands to be potentially disastrous in the case of norms allowing for social cooperation. Instead of reading him as saying 'cheat if you like, and here's a good way to get away with it!' then, we might read him as saying 'if you must cheat, or already have done, you ought at least to minimize the damage ... and here's how'. The risk is that too many people cheat-and-minimize-the-damage, which places the norm at risk. In the next section, we will look in more detail at that risk.

3. Goodin-style Cheaters in the Real World

To what extent should we expect to find Goodin-style cheaters in the real world? Let's take another look at what is required of his defector:

> When honouring social norms in the breach, therefore, you must engage in the more florid and rhetorical displays associated with civil disobedience. You acknowledge the rule that you are breaking, and (*modulo* what I shall go on to say below) you openly acknowledge that you are indeed breaking it. You engage in lots of hand-wringing, you go on and on about how hard the decision was, how very atypical were the circumstances in which you found yourself. You promise to faithfully comply with the norm under other circumstances in the future, and you entreat others to do likewise. You emphasize that your action should not be taken as a precedent by others. (296)

Goodin's would-be norm-breaker must publicly acknowledge that he's breaking the norm (presumably because protestations are not so convincing if caught trying to get away with something), must appear visibly distressed, must make convincing promises about the future, must persuade others not to follow his defection. He must have the charm to pull all of this off without arousing suspicion. In terms of acknowledging the rule that is being broken, Goodin suggests five ways of escaping the obligation. a) Say that the norm people follow is not the 'true' norm, while the one you followed is, perhaps appealing to God as the Author of the true norms. The problem with this escape is that if others believe you, they're likely to want to follow the true norm too, thus triggering the normative cascade, and if they don't believe you, they'll take you as defecting, which opens up the possibility of their like defection (but doesn't necessitate it; they might think you made an honest mistake). Either way, there's a risk of norm collapse. b) Say you're adhering to the principles that underly norms, perhaps principles about social cohesion. But the response to this will be just the same as the response just given. c) Claim weakness of will, saying that the norm is a good one but you couldn't bring yourself to obey it. The problem with this excuse is that there's no reason why others shouldn't use it, and it won't fly in cases where the duty wasn't very difficult. d) Claim overdemandingness, saying that the norm is a good one but it only applies to people who can fulfill it, and you can't. This is similar to e) claim that you're an exception to the rule, but that you don't advocate rule change to allow for exceptions because you believe that rules should be general in form.

Both (d) and (e) raise the question of whether obligations are conditional or unconditional in form. If they are unconditional, then neither excuse will work. 'Impossible' isn't the same as 'really hard', and the response will just be that you're required to do your duty (e.g. obey the norm) whether you find it difficult or not; overdemandingness should be reserved for cases where a duty is strictly impossible for an agent to fulfill. And claiming to be an exception won't work either, because if the rule was meant to allow for exceptions you might expect people to generally act in a way consistent with there being exceptions, and not be angered by defections that obviously occur in special circumstances (although perhaps the real difficulty here is the opacity of people's reasons for action). But it can't be that *all* of our obligations are conditional instead. If that's true, then it means that only if conditions x and y are satisfied, do I have an obligation to do z. If a person is robbed and I am the only person around, then I have an obligation to help them. But that creates two problems.

Firstly, it would commit us to thinking that there are no unconditional obligations, for instance to help people in need. Secondly, it would seriously violate Occam's Razor and clutter up our moral ontology. Rather than a limited set of unconditional

principles covering infinite actual cases, we'd find ourselves with an infinite set of complicated conditional requirements, designed to cover all possible eventualities and say what is required of people under each of them. Surely even if most of our obligations are conditional, there are some unconditional requirements, or there is a way of expressing the conditional requirements in a way that makes them general (perhaps we can do without a conditional operator and express them as primitives). If our obligations do work in this way, then (d) and (e) stand a chance of succeeding. Then the argument would be that the norm in question applies to cases meeting criteria *p* and *q*, which are met for most people most of the time, but not for cases meeting criteria *n*—which may well be that fulfilling the norm would be too difficult—which you yourself happen to meet. The obvious limit on this kind of defection is that there just won't be that many plausible excuses, exceptions which other people don't also fall into. People who find themselves in a similar situation can either take themselves to be excused from obeying the norm, or take you to not be. This is a good thing so far as society is concerned, because it is likely that it can absorb the rare impact of norm-breaking in exceptional circumstances. It is less good news for the social cheater concerned to exploit Goodin's conditions to his own advantage, because it means opportunities will be extremely infrequent.

In *Passions within Reason*, Robert Frank argues that certain signs of reliability, such as those Goodin's cheater would have to perform, are hard to fake. He introduces three principles of signaling. Firstly, for something to function as a reliable signal, it must be difficult (or costly) to fake. The example he uses is the reliable correlation in toads between their size and the depth of their croak, and how physical limitations simply prevent a very small toad from emitting a very deep croak. Secondly, having some characteristic must have benefitted the first individual who had it, and this will usually mean that signals evolved for reasons unrelated to the signaling benefit they came to have. Thirdly, the fact that someone signals positive information forces others into full disclosure about their own quality, because if they remain silent they will be assumed to be worse that they actually are.

> The general message of this full-disclosure principle is that a lack of evidence that something resides in a favoured category will often suggest that it belongs in a less favoured one. (Frank 1988, ch. 5; see also Akerlof 1970)

Frank distinguishes passive from deliberate signals, arguing that all three conditions must hold in the case of the former, and only the first and last in the case of the latter. Because Goodin's cheater will be attempting a deliberate signal of trustworthiness and sincerity (breaking the norm but attempting to justify the breach) we should look more closely at how hard to fake that might be.

Frank runs through a whole host of physical events that occur to people under stress, and that happen under occurrence of genuine emotion but can be deliberately manipulated only by very few people. The underlying idea is that certain types of physical responses are instinctual (he uses the example of a dog preparing to fight, with its hair standing on end, teeth bared, back arched, and muscles tense), and that imitating them will be slow and difficult. There are what psychologists call 'reliable' facial muscles, which only a small portion of the population can control. For instance the downward pull of the mouth without movement of muscles in the chin can be achieved deliberately by only about 10% of the population, but we all make that movement automatically when experiencing grief (Frank 1988, ch. 6). Also difficult is the eyebrows raised and furrowed associated with sadness, grief and distress (which about 15% of people can control), the eyebrows raised and pulled together associated with fear and terror (which less than 10% can control), eliminating the fleeting micro-expressions which will cross the face before the deliberate expressions can replace them, holding a steady gaze, non-dilation of pupils, blinking at regular speed, normal levels of mouth moisture (the mouth dries under stress, which can show up in pronunciation), avoiding blushing, and managing regular modulation of voice, because experiments show that about 70% of subjects' voices go up in pitch when emotionally upset (Frank 1988, ch. 6; see also Ekman 1985).

Frank suggests that perhaps the best way for a cheater to cheat successfully, i.e. to avoid giving the game away with any of these manifold physical slips, is to deceive himself into thinking he really *is* innocent of cheating, or really *did* have a good reason that made him an exception to the social norm or rule. But the problem Frank points out is that self-deception must sacrifice guilt (if you believe you were right in cheating, then you can't simultaneously feel guilty for cheating). And if a person cannot express guilt, then they will be assumed not to feel it at all (by the full-disclosure principle).

A further complication Frank suggests for the cheater is that experimental evidence shows people to have a tendency toward discounting negative possibilities if they are in the distant future as opposed to the near future. This means that when confronted with an opportunity to cheat (even one with a high likelihood of getting caught) many people will be tempted to take the risk, knowing that they will receive *now* the short term benefit of whatever goods their cheating gains them. Frank suggests that a good reputation can nonetheless function as a reliable indicator of trustworthiness because the reliable person will have internalized a preference against cheating, in which case the potential guilt from doing so will sufficiently offset the short-term gains that cheating promises. Someone who is fine with cheating will have no such off-setting mechanism, and therefore be more likely to cheat, and so more likely to get caught, thus risking their reputation. Returning momentarily to the charitable and uncharitable readings of Goodin, the consequences of sincerity being hard to fake will be different for each. The cheater really is faking, so he is subject to the difficulties just discussed in a way the genuine exception (or the person who takes themselves to be a genuine exception) is not. And the harder sincerity is to fake, the more we can be sure that the only norm-breakers getting away with norm-breaking will be those who genuinely 'honour the norm in the breach'.

It has been fairly well-established that one way to stabilize human cooperation is to punish those who defect on a cooperative scheme. But punishment itself creates a public goods problem: punishing is costly to the individual punisher, but the benefit of punishing accrues to the whole group (although this does depend on the kind of sanctions involved. Alienating someone or restricting their liberty will be much more costly than merely expressing disesteem toward them, which might be cost

less). Costly signaling theory, the main rival of both strong and conditional reciprocity theories, suggests that the sanctioning of individuals who defect on their collective obligations, or free-ride upon others' contributions, can act as a costly signal of the signaler's quality. Fairness norms act as a test of people's willingness or ability to pay their fair share, thus identifying defectors as unreliable or undesirable partners in cooperative ventures. Smith and Bliege Bird (2005; see also 2000) argue that defection from a prosocial norm might signal one of three things, depending on the context: inability to pay one's share, which is a signal of low quality, leading to reduced social status; withdrawal, if the cooperation is a competition for status; or flaunting the norm to signal superior social status or power, a reliable interpretation so long as violating the norm is more costly than paying one's share (Smith/Bliege Bird 2005, 136). They suggest that the enforcement of fairness norms may be as much about ensuring the reliability of signals, and the solution of status-competition games, as it is about the importance of ensuring equality itself (137). Along similar lines, Joseph Bulbulia has argued that following religious norms is a costly signal of altruistic intentions, making the signaler a more attractive cooperation partner (Bulbulia 2004), and Samuel Bowles and Herbert Gintis have argued that certain kinds of prosocial cooperation, including the unconditional sharing of resources, participating in group defense, and punishing norm violation, is a costly signal of being a worthy mate, coalition partner, or competitor (Bowles/Gintis 2001). Neither of these latter two explicitly defends the idea that accepting the costs of punishing norm-breakers is an honest signal of quality, but to the extent that provision of punishment is just another public good which an elite few may provide to the whole group, the same arguments can be expected to apply.

One interesting issue to flag here is the distinction between merely obeying social norms that everyone else (or almost everyone) obeys, and obeying social norms that only an elite group can obey, e.g. where a few people provide a public good for the whole group. The idea is that costly signaling theory can explain how what looks like a puzzlingly altruistic set of actions, e.g. engaging in dangerous and physically demanding hunting to supply a communal feast (Smith/Bliege Bird 2000), or as suggested above, being the person willing to absorb the costs of punishing the non-cooperators, can be explained by virtue of being a costly signal of quality, which will have benefits for the signaler. The person who defects on the first order norms runs a double risk. The first, as Smith and Bliege Bird suggested, is signaling unreliability as a cooperative partner, by virtue of defection meaning in inability or an unwillingness to pay the cost associated with obeying the norm. The second is the risk of being sanctioned by the second-order norm follower looking to signal quality by punishing first-order defection. And it's not necessarily the case that a convincing justification of the norm breach will avoid those risks, because it might be the case that the association between norm-breaking and cooperative unreliability has become automatic or instinctual as a cognitive heuristic to calculating the likelihood of being cheated.

4. *Conclusion*

Goodin suggests that precisely the kinds of behaviours that allow conscientious objectors and international law reformers to break norms without gaining a reputation as a norm-breaker in the pejorative sense could be emulated in the domestic case to allow successful defection from social norms. In this paper I have asked whether the kind of well-contained norm breaking he suggests is actually worth worrying about. I argued that it is, because of the burden it places on others in a context where defection creates a remainder. I have also asked whether Goodin-style defectors are likely to be common in the real world, and I argued, for various reasons stemming from the difficulty of faking sincerity of the kind required, and the risk that defection poses to the individual defector, that we can expect such cheating to be fairly rare.

Bibliography

Akerlof, G. (1970), The Market for Lemons, in: *Quarterly Journal of Economics 84*, 488–500

Axelrod, R. (1984), *The Evolution of Cooperation*, New York

— (1986), An Evolutionary Approach to Norms, in: *American Political Science Review 80*, 1095–1111

Bulbulia, J. (2004), Religious Costs as Adaptations that Signal Altruistic Intentions, in: *Evolution and Cognition 10(1)*, 19–42

Ekman, P. (1985), *Telling Lies*, New York

Frank, R. (1988), *Passions within Reason*, New York

Gintis, H./E. A. Smith/S. Bowles (2001), Costly Signalling and Cooperation, in: *Journal of Theoretical Biology 213(1)*, 103–119

Goodin, R. (2005), Toward an International Rule of Law: Distinguishing International Law-Breakers from Would-Be Law Makers, in: *The Journal of Ethics 9*, 225–246

Pettit, P. (2008), Norms, Commitment and Censure, paper presented at colloquium *Hart/Fuller 50 Years On*, Australian National University, Canberra, December 2008

Smith, E. A./R. Bliege Bird (2005), Costly Signalling and Co-operative Behaviour, in: H. Gintis/S. Bowles/R. Boyd/E. Fehr (eds.), *Moral Sentiments and Material Interests*, Massachusetts, 115–150

— / — (2000), Turtle Hunting and Tombstone Opening: Public Generosity as Costly Signalling, in: *Evolution and Human Behaviour 21*, 245–261