Behavioural Public Policy (2019), 3: 1, 1–16
 doi:10.1017/bpp.2018.8

Growing outrage

CASS R. SUNSTEIN

Robert Walmsley University Professor, Harvard University, Boston, MA, USA

Abstract: Why and when does outrage grow? This essay, based on the 2018 LSE Behavioural Public Policy Lecture delivered in February 2018, explores two potential answers. The first points to a revision or weakening of social norms, which leads people to express outrage that they had previously suppressed. The second points to a revision or weakening of social norms, which leads people to express outrage that they had not previously felt (and may or may not now feel). The intensity of outrage is often a product of what is most salient. It is also a product of 'normalization'; people compare apparently outrageous behavior to behavior falling into the same category in which it is observed, and do not compare it to other cases, which leads to predictable incoherence in judgments. These points bear on the #MeToo movement of 2017 and 2018 and the rise and fall (and rise again, and fall again) of discrimination on the basis of sex and race (and also religion and ethnicity).

Submitted 14 February 2018; accepted 14 February 2018

I begin with two tales.

1. In the late 1980s, when I was a visiting professor at Columbia Law School, I happened to pass, in the hallway near my office, a law student (female) speaking to an older law professor (male). To my amazement, the professor was stroking the student's hair. I thought I saw, very briefly, a grimace on her face. It was a quick flash. When he left, I said to her, "That was completely inappropriate. He shouldn't have done that." Her response was immediate and dismissive: "It's fine. He's an old man. It's really not a problem." Thirty minutes later, I heard a knock on my door. It was the student. She was angry, and she was outraged, and she was in tears. She said, "He does this all the time. It's horrible. My boyfriend thinks I should make a formal

Email: csunstei@law.harvard.edu

complaint, but I don't want to do that. Please – I don't want to make a fuss. Do not talk to him about it and do not tell anyone."[1]

2. In 2015, I was walking through a cafeteria at my law school when a group of African–American students stopped me to ask, "What do you think of the seal?" I had no idea what they were asking about. My immediate thought was that something bad might have happened to a seal – say, at the New England Aquarium – and so I almost asked, "Is the seal ok?" But they were asking about the Harvard Law School seal, which had come from a slave-owner. The seal became the source of considerable local outrage and spurred student protests on a wide variety of race-related issues. It was eliminated in 2016.

My topic here is outrage, its growth, and its relationship to the idea of #MeToo, writ very large. In 2017, that particular hashtag went viral, spurred as it was by disclosures of sexual assault and related behavior by Harvey Weinstein, and also of sexual harassment by a wide range of public figures. Both of the tales just told have #MeToo features, but the dynamics are different. I am acutely aware of the sheer size of the topic, and will mostly restrict myself to two propositions, which the tales respectively exemplify.

The first is that that when norms start to collapse or to be revised, people are unleashed, in the sense that *they feel free to reveal what they believe and prefer, to disclose their experiences, and to talk and act as they wish*. If they are outraged, they can disclose that fact. New norms, and laws that entrench or fortify them, lead to the discovery of pre-existing beliefs, preferences, and values. Discovery of the intensity of people's outrage, and even its existence, can be startling, at least to many people.

In various times and places, the women's movement has been an example. The same is true for the civil rights movement of the 1960s, the movement for LGBT rights, and the disability rights movement. It is also true for the pro-life movement. Often an antecedent feeling of humiliation is the foundation for outrage. Once norms are relaxed, outrage can become an immensely powerful force.

The second proposition, exemplified by the second tale, is that revisions of norms can construct preferences and values, leading to an expression of outrage that did not exist before. No one is unleashed. People are changed in ways large or small. They express outrage that they may not actually feel – and if they feel it, it is for the first time. Something like this can be said for

1 What I did, in response to this exchange, is a story for another occasion.

the anti-smoking movement, the objections to genetically modified organisms, and the rise of Nazism.

For new or revised norms to be effective, of course, they usually have to be able to attach themselves to pre-existing preferences or values; that is certainly true in my second tale. Outrage about the Harvard Law School seal was connected with pre-existing judgments about race and racism. The only point is that people come to make judgments (whether particular or general) that they did not hold before, and that they had no occasion to suppress. (Before 2014, hardly anyone was exercised about the Harvard Law School seal.)

With respect to the phenomenon of unleashing: when certain norms are in force, people falsify their preferences, their values, and their experiences, or are silent about them. As a result, strangers and even friends and family members may not be able to know about them. They may have no idea. People with certain political or religious convictions, or simply painful experiences, might just shut up, even if they are outraged. Once norms are revised, people will reveal pre-existing preferences and values, which norms had successfully suppressed. What was once unsayable is said, and what was once unthinkable is done.

In the context of sexual harassment, something like this account is broadly correct: before the term itself existed, women did not like being harassed, or even hated it, and revision of social norms was necessary to spur expression of their feelings and beliefs. (This account is incomplete, and I will complicate it.) As we shall see, law often plays a significant role in fortifying existing norms, or in spurring their revision. Part of the importance of judicial rulings that forbid sexual harassment is that they revised norms. Whenever a new leader is elected, or whenever new legislation is enacted, it may have a crucial and even transformative signaling effect, offering people information about what other people think. If people hear the signal, norms may shift, because people are influenced by what they think other people think. Outrage is often fueled in that way.

But some revisions of norms, and the laws that entrench those revisions, do not liberate anything. As norms begin to be altered, people come to hold, or to act as if they hold, preferences and values that they did not hold before. Revisions of norms, and resulting legal reforms, do not uncover suppressed desires. They produce new ones, or at least statements and actions that are consistent with new ones. Outrage can fill a vacuum.

Preference falsification and norm entrepreneurs

Jon Elster emphasizes that social norms are "sustained by the feelings of embarrassment, anxiety, guilt, and shame that a person suffers at the prospect of

violating them" (Elster, 1994, p. 23). Elster's quartet is worth underlining: embarrassment, anxiety, guilt, and shame are different from one another. The student at Columbia Law School felt all four.

My central topic here is discrimination. In the simplest cases, the objects of discrimination (emphatically including sexual harassment) are outraged, and the norm prevents them from stating or acting on it. Their beliefs and preferences may even be falsified (as it was when the law student initially assured me that she did not object to what the professor was doing). In that respect, the objects of discrimination are like actors in a play; they are reciting the expected lines. In cases of sex and race discrimination, that is a familiar phenomenon. The legitimation of outrage brings it out of the closet.

In circumstances of this kind, large-scale change is possible. Suppose that many people within a population object to discrimination, but because of existing norms, they do not say or do anything. They falsify their preferences.[2] Suppose that the objectors have different thresholds for raising an objection. A few people will do so if even one person challenges or defies the norm; a few more will do so if a few people challenge or defy the norm; still more will do so if more than a few people challenge or defy the norm; and so on. Under the right conditions, and with the right distribution of thresholds, a small spark can ignite a conflagration, eventually dismantling the norm.[3]

There is an important role here for 'norm entrepreneurs', operating in the private or public sector, who oppose existing norms and try to change them. Norm entrepreneurs draw attention to what they see as the stupidity, intrusiveness, or ugliness of current norms. They may insist that many or most people secretly oppose them (and thus reduce pluralistic ignorance, understood as ignorance about what most people actually think). They may describe their experiences. Norm breakers – those who simply depart from existing norms, and refuse to speak or act in accordance with them – may or may not be norm entrepreneurs, depending on whether they seek to produce some kind of social change, or instead wish merely to do as they like.

Norm entrepreneurs might turn out to be effective, at least if the social dynamics work out in their favor. They might be able to signal not only their outrage and personal opposition to the norm, but also the existence of widespread (but hidden) outrage and opposition as well. The idea of a 'silent majority' can be a helpfully precise way to signal such outrage and opposition. Importantly, norm entrepreneurs might also change the social meaning of

2 The best discussion of the general phenomenon and its importance is Timur Kuran (1997).

3 The classic account is Mark Granovetter (1978); the idea is productively extended in Timur Kuran (1997).

compliance with the norm: if they succeed, such compliance might suggest a lack of independence and look a bit pathetic, whereas those who defy the norm might seem courageous, authentic, and tough.

The outrage heuristic

A number of years ago, I was involved in a series of studies of outrage, punitive intentions, and monetary punishments. Our basic finding was that when ordinary people are thinking about how much to punish people, they use the *outrage heuristic*.[4] They begin by deciding how outrageous the underlying conduct was, and their judgments about punishment build on that decision. We found that people's outrage judgments, on a bounded numerical scale, almost exactly predicted their punitive intentions on the same scale. That means that people are *intuitive retributivists*. Unless prompted, they do not think about optimal deterrence (and even when prompted, they resist the idea).

One of our studies tested the effects of deliberation on both punitive intentions and monetary judgments (Schkade *et al.*, 2000). The study involved about 3000 jury-eligible citizens; its major purpose was to determine how individuals would be influenced by seeing and discussing the punitive intentions of others. Our central goal was to explore how social interactions heighten outrage.

People were initially asked to record their individual judgments privately, on a bounded scale, and then to join six-member groups to generate unanimous 'punishment verdicts'. Hence, subjects were asked to record, in advance of deliberation, a 'punishment judgment' on a scale of 0–8, where 0 indicated that the defendant should not be punished at all and 8 indicated that the defendant should be punished extremely severely. (Recall that outrage judgments on such scales are mirrored by punishment judgments, so we were essentially measuring outrage.) After the individual judgments were recorded, jurors were asked to deliberate to a unanimous 'punishment verdict'. It would be reasonable to predict that the verdicts of juries would be the median of punishment judgments of jurors, but that prediction would be badly wrong.

The finding that I want to emphasize here is that deliberation made the lower punishment ratings *decrease* when compared to the median of pre-deliberation judgments of individual jurors – while deliberation made the higher punishment ratings *increase* when compared to that same median. When the individual jurors favored little punishment, the group showed a 'leniency shift',

4 Most of the work was done in collaboration with Daniel Kahneman and David Schkade. For a collection, see Cass R. Sunstein *et al.* (2007).

meaning a rating that was systematically lower than the median pre-deliberation rating of individual members. That means that when people began with low levels of outrage, deliberation produced lower levels still. But when individual jurors favored strong punishment, the group as a whole produced a 'severity shift', meaning a rating that was systematically higher than the median pre-deliberation rating of individual members. In groups, outrage grows.

Outrage and group polarization

What accounts for the leniency shift and the severity shift?

The simplest answer lies in the phenomenon of group polarization (see Sunstein, 2009). This is the pervasive process by which group members end up in a more extreme position in line with the pre-deliberation tendencies of group members. It is now well known that if a group has a defined median position, members will shift toward a more extreme version of what they already think. Consider some examples of the basic phenomenon, which has been found in over a dozen nations (see Brown, 1985, p. 222)[5]: (a) a group of moderately pro-feminist women will become more strongly pro-feminist after discussion (see Myers, 1975); (b) after discussion, citizens of France become more critical of the United States and its intentions with respect to economic aid (Brown, 1985, p. 224); (c) after discussion, whites predisposed to show racial prejudice offer more negative responses to the question of whether white racism is responsible for conditions faced by African–Americans in American cities (Myers & Bishop, 1976); and (d) after discussion, whites predisposed not to show racial prejudice offer more positive responses to the same question (Myers & Bishop, 1976).

Why does deliberation drive low punishment ratings down and move high punishment ratings up? There are three answers. The first involves the exchange of information within the group. In a group that favors a high punishment rating, group members will make many arguments in that direction and relatively few the other way. Speaking purely descriptively, the group's 'argument pool' will be skewed in the direction of severity. Group members, listening to the various arguments, will naturally move in that direction. The initial dispositions of group members will determine the proportion of arguments in the various directions. And individuals will respond, quite rationally,

5 These include the United States, Canada, New Zealand, India, Bangladesh, Germany, and France (see, e.g., Zuber *et al.*, 1992 [Germany]; Abrams *et al.*, 1990, p. 112 [New Zealand]). Of course, it is possible that some cultures would show a greater or lesser tendency toward polarization; this would be an extremely interesting area for empirical study.

to what they have heard, thus moving in the direction suggested by the dominant tendency. In this way, outrage breeds more outrage.

The second explanation involves social influences. Most people want to be a certain way and also to be perceived in a certain way. If you are in a group that is outraged and wants to punish someone severely, you might find it uncomfortable to be urging relative leniency. To protect your reputation, and perhaps your self-conception, you might move, if you move at all, in the most favored direction. To be sure, some hardy souls will not move at all, and those who are self-identified contrarians might deliberately move in the opposite direction, rejecting the dominant view just because it is the dominant view. But what we observed, and what is generally observed, is that most of those who move tend to go in the group's preferred direction – and that as a result, the group will be more extreme than its members before deliberation began. With respect to outrage, the lesson is clear. To preserve their preferred self-image, individuals, finding themselves in an outraged group, will tend to become more outraged still.

The third explanation begins by noting that people with extreme views tend to have more confidence that they are right and that as people gain confidence, they become more extreme in their beliefs. The intuition here is simple: those who lack confidence, and who are unsure what they should think, tend to moderate their views. It is for this reason that cautious people, not knowing what to do, are likely to choose the midpoint between relevant extremes. But if other people seem to share one's view, that person is likely to become more confident that that view is right – and hence to move in a more extreme direction.

We can easily see how the third explanation might apply to outrage in particular. Someone has done something that seems wrong. It might be a spouse, who has spoken unkindly, rudely, or cruelly. It might be an employer, who has exceeded the appropriate bounds in one or another way. It might be a corporation. It might be a public official. If people are asked about their reactions in their purely individual capacities, they might think: 'not good'. But if they are speaking with one another, they might end up confirming one another's initial instincts, leading to greater confidence and eventually to the thought: 'horrific; intolerable'. A supplemental point involves salience. Discussion of bad conduct will heighten people's attention to it, leading to more intense reactions. What was once a background fact, or part of life's furniture, might become one of the most important things in the world.

Normalization and categories

Levels of outrage are specific to categories. If someone is very rude on Twitter, at a lunch table, on the highways, in a security line, in a meeting, or in

comments on an academic paper, people might see red; it is as if something truly awful has been done. By contrast, some actual crimes (say, shoplifting or in some cases tax evasion) might produce only a modest level of outrage. But on reflection, people would agree that rudeness is less outrageous than criminality (or at least most forms of criminality). In the context of outrage and punishment decisions, our most striking finding is that *people's judgments about cases, taken one at a time, are very different from their judgments about the very same cases, taken in the context of a problem from another category* (see Sunstein, 2002).

An example: people were asked to assess a case involving personal injury on a bounded punishment scale and also on a monetary scale. People were also asked to assess a case involving financial injury, again on a bounded punishment scale and on a monetary scale. When the two cases are judged in isolation, the financial injury case receives a more severe rating and a higher monetary award. But when the two cases are seen together, there is a significant 'judgment shift', in which people ensure that the financial award is not much higher, and for many respondents is lower, than the personal injury award. In short, people's decisions about the two cases are very different, depending on whether they see the case alone or in the context of a case from another category.

Notice that monetary awards shift, and that outrage (the foundation of intention to punish) shifts as well. Apparently the level of outrage will differ depending on whether a case is seen in isolation or instead in the context of cases from other categories.

Exactly the same kind of shift is observed for judgments about two problems calling for government regulation and expenditures: skin cancer among the elderly and protection of coral reefs. Looking at the two cases in isolation, people will pay more to protect coral reefs, and register more satisfaction from doing that. But looking at the two cases together, people will be quite disturbed at this pattern, and will generally want to pay more to protect elderly people from cancer. Here, too, there is a significant shift in judgment.

Is this a problem? And what accounts for the switch? Consider a preliminary account. When people see a case in isolation, they naturally 'normalize' it by comparing it to a set of comparison cases that it readily calls up. If you are asked, 'Is a Great Dane big or small?', you are likely to respond that it is big; if you are asked, 'Is a Toyota Tercel big or small?', you are likely to respond that it is small. But people are well aware that a Great Dane is smaller than a Toyota Tercel. People answer as they do because a Great Dane is compared with dogs, whereas a Toyota Tercel is compared with cars. So far, so good; in these cases, everyone knows what everyone else

means. We easily normalize judgments about size, and the normalization is mutually understood.

In the context of outrage, something similar happens, but it is less innocuous. In general, the level of outrage undergoes a similar process of normalization. If an academic colleague has offended you, for example by saying that your recent work is 'dreadful' and 'should not have been published', you might be extremely outraged, simply because you compare the comments to ordinary collegial behavior, and do not naturally compare them to other sorts of behavior, such as rape and assault. It might take a self-conscious cognitive exercise to decide that the offensive comments are not, in the scheme of things, deserving of a high level of outrage.

When evaluating a case involving financial injury, people apparently 'normalize' the defendant's conduct by comparing it with conduct in other cases *from the same category*. They do not easily or naturally compare that defendant's conduct with conduct from other categories. Because of the natural comparison set, people are likely to be quite outraged by the misconduct, if it is far worse than what springs naturally to mind. The same kind of thing happens with the problem of skin cancer among the elderly. People compare that problem with other similar problems – and conclude that it is not so serious, within the category of health-related or cancer-related problems. So too with personal injury cases (normalized against other personal injury cases) and problems involving damage to coral reefs (normalized against other cases of ecological harm).

The key point is that when a case from another category is introduced, this natural process of comparison is disrupted. Rather than comparing a skin cancer case with other cancers, or other human health risks, people see that it must be compared with ecological problems, which (in most people's view) have a lesser claim to public resources. Rather than comparing a financial injury case to other cases of business misconduct, people now compare it to a personal injury case, which (in most people's view) involves more serious wrongdoing. As a result of the wider view screen, judgments of outrageousness and appropriate punishment shift, often dramatically.

Most of the time, people's failure to use a wide view screen, in thinking about the appropriate degree of outrage, is not damaging. That failure is a way of economizing on thinking. But for law and policy, the process of normalization, and the use of a narrow view screen, produces serious problems. The difficulty is that when people assess cases in isolation, their view screen is narrow, indeed limited to the category to which the case belongs, and that as a result, people produce a pattern of outcomes that makes no sense by their own lights. In other words, the overall set of outcomes is one that people would not

endorse, if they were only to see it as a whole. (With respect to outrage, readers can think of their own preferred examples.)

Parallel worlds and multiple equilibria

It is important to emphasize that with small variations in starting points, and inertia, resistance, or participation at the crucial points, significant changes in statements or in actions may or may not happen. Outrage may fizzle or grow.

Suppose that a community has long had a norm in favor of discrimination on the basis of sexual orientation; that many people in the community abhor that norm; that many others dislike it; that many others do not care about it; that many others are mildly inclined to favor it; and that many others firmly believe in it. If norm entrepreneurs make a public demonstration of opposition to the norm, and if the demonstration reaches those with relatively low thresholds for opposing it, opposition will immediately grow. If the growing opposition reaches those with relatively higher thresholds, the norm might rapidly collapse. In many places in the world, that is exactly what happened in recent decades. But if the early public opposition is barely visible, or if it reaches only those with relatively high thresholds, it will fizzle out, and the norm might not even budge. In many places in the world, that has happened, too.

These are the two extreme cases. We could easily imagine intermediate cases, in which the norm suffers a slow, steady death, or in which the norm erodes but manages to survive. It is for this reason that otherwise similar communities can have multiple equilibria, understood here as apparently or actual stable situations governed by radically different norms. After the fact, it is tempting to think that because of those different norms, the communities are not otherwise similar at all, and to insist on some fundamental cultural difference between them. But that thought might well be a product of an illusion, in the form of a failure to see that some small social influence, shock, or random event was responsible for the persistence of a norm in one community and its disintegration in another.

Some of the most interesting work on social influences involves the existence of informational and reputational 'cascades'; this work has obvious relevance to the growth of outrage (see Bikhshandani *et al.*, 1992). A starting point is that when individuals lack a great deal of private information (and sometimes even when they have such information), they are attentive to the information provided by the statements or actions of others. If A is unaware whether genetic modification of food is a serious problem, he may be moved in the direction of alarm if B seems to think that alarm is justified. If A and B believe that

alarm is justified, C may end up thinking so too, at least if she lacks independent information to the contrary. If A, B, and C believe that genetic modification of food is a serious problem, D will have to have a good deal of confidence to reject their shared conclusion. The result of this process can be to produce cascade effects, as large groups of people eventually end up believing something simply because other people seem to believe it too. It should be clear that cascade effects may occur, or not, depending on seemingly small factors, such as the initial distribution of beliefs, the order in which people announce what they think, and people's thresholds for abandoning their private beliefs in deference to the views announced by others.

Though the cascades phenomenon has been discussed largely in connection with factual judgments, the same processes are at work for norms; we can easily imagine outrage cascades (information-induced or otherwise), which may well produce social change and legal reform.[6] Some such cascades may be a product of information; some may involve values. Suppose, for example, that A believes that discrimination against transgender people is wrong, that B is otherwise in equipoise but shifts upon hearing what A believes, that C is unwilling to persist in his modest approval of discrimination against transgender people when A and B disagree; it would be a very confident D who would reject the moral judgments of three (apparently) firmly committed others. In such contexts, many people, lacking firm convictions of their own, may end up believing what (relevant) others seem to believe.

Stylized as the example is, changes in social attitudes toward smoking, recycling, and sexual harassment have a great deal to do with these effects. And here as well, small differences in initial conditions, in thresholds for abandoning private beliefs because of reputational pressures, and in who hears what when, can lead to major differences in outcomes. And again: after the fact, it may all seem inevitable, a product of historical forces, even if serendipity played an essential role.

A 'down look'

For discrimination, of course, it is too simple to say that its objects are opposed and suppress their outrage. When discrimination is widespread, and when existing norms support it, its objects might see it as part of life's furniture.

6 An intriguing wrinkle is that when a cascade gets going, people might underrate the extent to which those who join it are reacting to the signals of others, and not their own private signals. For that reason, they might see the cascade as containing far more informational content than it actually does (see Eyster & Rabin, 2010; Eyster *et al*., 2015). Norm entrepreneurs have a strong interest in promoting this mistake.

That metaphor buries a lot of complexity. Sometimes people may feel hurt or burdened, but their sense of pain or injury may not be transformed into a claim or even a feeling of injustice. Some preferences are *adaptive*; they are a product of existing injustice. If a victim of sexual harassment genuinely believes that 'it's not a big deal', it might be because it's most comfortable or easiest to believe that it's not a big deal. There is a spectrum here, from real pain (but still, it's not a big deal) to a feeling that it is just life (and not really painful).

Consider Gordon Wood's account of the pre-revolutionary American colonies, when "common people" were "made to recognize and feel their subordination to gentlemen," so that those "in lowly stations … developed what was called a 'down look'," and "knew their place and willingly walked while gentlefolk rode; and as yet they seldom expressed any burning desire to change places with their betters" (Wood, 1998, pp. 29–30). In Wood's account, it is impossible to "comprehend the distinctiveness of that premodern world until we appreciate the extent to which many ordinary people *still accepted their own lowliness*" (Wood, 1998, pp. 29–30, emphasis added). (Is Wood right? Did they really accept their own lowliness? It's hard to know – but let's bracket that point.)

Wood urges that as republicanism took hold, social norms changed, and people stopped accepting their own lowliness. His account is one of an outrage cascade, but not as a result of the revelation of pre-existing preferences. With amazement, John Adams wrote that "Idolatry to Monarchs, and servility to Aristocratical Pride, was never so totally eradicated from so many Minds in so short a Time" (Wood, 1998, p. 169). David Ramsay, one of the nation's first historians (himself captured by the British during the American Revolution), marveled that Americans were transformed "from subjects to citizens," and that was an "immense" difference, because citizens "possess sovereignty. Subjects look up to a master, but citizens are so far equal, that none have hereditary rights superior to others" (Wood, 1998, p. 169). Thomas Paine put it this way: "Our style and manner of thinking have undergone a revolution more extraordinary than the political revolution of a country. We see with other eyes; we hear with other ears; and think with other thoughts, than those we formerly used" (Paine, 1908, p. 242).

Adams, Ramsay, and Paine appear to be speaking of new preferences, beliefs, and values, rather than the revelation of suppressed ones. In their account, nothing is unleashed. How new preferences arise remains imperfectly understood. While the framework of preference falsification and unleashing captures much of the territory I am exploring, it is complemented by situations in which adaptive preferences are altered by new or revised norms.

There are also intermediate cases, involving what might be called *partially adaptive preferences*. Objects of discrimination may not exactly accept

discrimination. As noted, they might feel pain or a burden. They might live with it, and do so with a degree of equanimity, thinking that nothing can be done. It is not a lot of fun to beat your head against the wall. In cases of partially adaptive preferences, objects of discrimination are not like actors in a play; they are not falsifying their preferences. They hear a small voice in their heads. They might even feel outrage, but it simmers. Once norms change, some inchoate belief or value might be activated that was formerly suppressed, or that was like that small voice. It is fair enough to speak of liberation, but the case is not as simple as that of the law student at Columbia.

Liberating outrage

Some norms reduce discrimination, but others increase it. Suppose that people have antecedent hostility toward members of social groups; suppose that social norms constrain them from speaking or acting in ways that reflect that hostility. On one view, this is the good side of 'political correctness'; it prevents people from expressing ugly impulses. But norms that constrain sexism and racism are of course stronger in some times and places than in others, and they can be relaxed or eliminated. In the aftermath of the election of President Donald Trump, many people fear that something of this kind has happened (and are fearing that it continues to happen). The basic idea is that President Trump is a norm entrepreneur; he is shifting norms in such a way as to weaken or eliminate their constraining effects. He is allowing people to express their real concerns, including their sense of outrage. It is difficult to test that proposition in a rigorous way, but consider a highly suggestive experiment.

Leonardo Bursztyn of the University of Chicago, Georgy Egorov of Northwestern University, and Stefano Fiorin of the University of California at Los Angeles attempted to test whether President Trump's political success affects Americans' willingness to support, in public, a xenophobic organization (Bursztyn *et al.*, 2017). Two weeks before the election, Bursztyn and his colleagues recruited 458 people from eight states that the website Predictwise said that Trump was certain to win (Alabama, Arkansas, Idaho, Nebraska, Oklahoma, Mississippi, West Virginia, and Wyoming). Half the participants were told that Trump would win. The other half received no information about Trump's projected victory.

All participants were then asked an assortment of questions, including whether they would authorize the researchers to donate $1 to The Federation for American Immigration Reform, accurately described as an anti-immigrant organization whose founder has written, "I've come to the point of view that for European–American society and culture to persist

requires a European–American majority, and a clear one at that" (Bursztyn *et al.*, 2017, p. 14). If participants agreed to authorize the donation, they were told that they would be paid an additional $1. Half the participants were assured that their decision to authorize a donation would be anonymous. The other half were given no such assurance. On the contrary, they were told that members of the research team might contact them, thus suggesting that their willingness to authorize the donation could become public.

For those who were not informed about Trump's expected victory in their state, giving to the anti-immigration group was far more attractive when anonymity was assured: 54% authorized the donation under cover of secrecy as opposed to 34% when the authorization might become public. But for those who were informed that Trump would likely win, anonymity did not matter at all. When so informed, about half the participants were willing to authorize the donation regardless of whether they received a promise of anonymity (Bursztyn *et al.*, 2017). The central point is that information about Trump's expected victory altered social norms, making many people far more willing to give publicly and eliminating the comparatively greater popularity of anonymous endorsements.

As an additional test, Bursztyn and his colleagues repeated their experiment in the same states during the first week after Trump's election. They found that Trump's victory also eliminated the effects of anonymity – again, about half the participants authorized the donation regardless of whether the authorization would be public. The general conclusion is that if Trump had not come on the scene, many Americans would refuse to authorize a donation to an anti-immigrant organization unless they were promised anonymity. But with Trump as president, people feel liberated. Anonymity no longer matters, apparently because Trump's election has weakened the social norm against supporting anti-immigrant groups. It is now more acceptable to be known to agree "that for European–American society and culture to persist requires a European–American majority, and a clear one at that" (Bursztyn *et al.*, 2017, p. 14).

The central finding can be seen as the mirror image of the tale of the law student and the law professor. For a certain number of people, hostility to anti-immigrant groups is a private matter; they do not want to voice that hostility in public. But if norms are seen to be weakening or to be shifting, they will be willing to give voice to their beliefs.

We can easily imagine much uglier versions of the central finding. When police brutality increases, when hateful comments or actions are directed at members of certain religious groups, when white supremacy marches start, when ethnic violence breaks out, when mass atrocities occur, and when genocide is threatened, one reason is the weakening or transformation of the social norms that once made the relevant actions unthinkable.

Internalized norms

My emphasis has been on situations in which people have an antecedent sense of outrage, whose expression a norm blocks; revision of the norm liberates them, so that they can talk or act as they wish. I have also noted that some norms are internalized, so that people do not feel chained at all. Once the norm is revised, they speak or act differently, perhaps expressing outrage, either because they feel constrained by the new norm to do that, or because their preferences and values have actually changed. Orwell's *Nineteen Eighty-Four* is a chilling tale of something like that, with its terrifying closing lines: "But it was all right, everything was all right, the struggle was finished. He had won the victory over himself. He loved Big Brother" (Orwell, 1949, p. 289).

That is the dark side. But return to the case of sexual harassment. Many men are appalled by the very thought of sexual harassment. They endorse, and do not feel constrained by, norms against it. For them, norms and legal rules against sexual harassment are not a problem, any more than norms and legal rules against theft and assault are a problem. For such men, we do not have cases of preference falsification. For some of them, it might be clarifying to speak of adaptive preferences. But it is better to say that the relevant people are committed to the norm, so that defying it would not merely be costly; it would be unthinkable.

Something similar can be said for many actions that conform to social norms. Most people are not outraged by the nonexistence of a social norm against dueling. For many people, seatbelt buckling and recycling are not properly characterized as costs; they are a matter of routine, and for those who buckle their seatbelts or recycle, the relevant actions may well be taken as a net benefit. When the social norm is one of considerateness (see Ullmann-Margalit, 2017), those who are considerate usually do not feel themselves to be shackled; they want to be considerate. When this is so, the situation will be stable; norm entrepreneurs cannot point to widespread, but hidden, outrage or dissatisfaction with the norm. But for both insiders and outsiders, it will often be difficult to distinguish between situations in which norms are internalized and situations in which they merely seem to be. That is one reason that stunning surprises are inevitable.[7]

7 Recall that another reason for unpredictability involves interdependencies among agents, which can produce changes that cannot be anticipated in advance (Lohmann, 2000).

Acknowledgments

This essay draws on, and can be seen as a companion piece to, Cass R. Sunstein, Unleashed, *Social Research* (forthcoming 2018). I am grateful to the editors of *Social Research* for permission to draw on that essay here.

References

Abrams, D. *et al.* (1990), 'Knowing What To Think By Knowing Who You Are', *British Journal of Social Psychology*, **29**(2): 97–119.

Bikhshandani, S., D. Hirshleifer, and I. Welch (1992), 'A Theory of Fads, Fashion, Custom, and Cultural Change As Informational Cascades', *Journal of Political Economy*, **100**(5): 992–1096.

Brown, R. (1985), *Social Psychology*, 2nd ed. New York: Free Press.

Bursztyn, L., G. Egorov and S. Fiorin (2017), From Extreme to Mainstream: How Social Norms Unravel. Available at: http://www.nber.org/papers/w23415 (Accessed: 15 February 2018).

Elster, J. (1994), 'Rationality, Emotions, and Social Norms', *Synthese*, **98**(1): 21–49.

Eyster, E. and M. Rabin (2010), 'Naïve Herding in Rich-Information Settings', *American Economic Journal: Microeconomics*, **2**(4): 221–243.

Eyster, E., M. Rabin and G. Weizsacker (2015), An Experiment on Social Mislearning. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2704746 (Accessed: 15 February 2018).

Granovetter, M. (1978), 'Threshold Models of Collective Behavior', *American Journal of Sociology*, **83**(6): 1420–1443.

Lohmann, S. (2000), 'I Know You Know He or She Knows We Know You Know They Know: Common Knowledge and the Unpredictability of Informational Cascades', In D. Richards (ed), *Political Complexity: Nonlinear Models of Politics*, Ann Arbor: University of Michigan Press.

Kuran, T. (1997), *Private Truths, Public Lies*, Cambridge: Harvard University Press.

Myers, D. G. (1975), 'Discussion-Induced Attitude Polarization', *Human Relations*, **28**(8): 699–714.

Myers, D. G. and G. D. Bishop (1976), 'The Enhancement of Dominant Attitudes in Group Discuission', *Journal of Personality and Social Psychology*, **20**(3): 386–391.

Orwell, G. (1949), *Nineteen Eighty-Four*, New York: Signet.

Paine, T. (1908), 'Letter to the Abbe Raynal', In D. E. Wheeler (ed), *Life and Writings of Thomas Paine*, New York: Vincent Parke and Company.

Schkade, D. *et al.* (2000), 'Deliberating About Dollars: The Severity Shift', *Columbia Law Review*, **100**(4): 1139–1175.

Sunstein, C. R. *et al.* (2002), 'Predictably Incoherent Judgments', *Stanford Law Review*, **54**(6): 1153–1215.

Sunstein, C. R. *et al.* (2007), *Punitive Damages: How Juries Decide*, Chicago: University of Chicago Press.

Sunstein, C. R. (2009), *Going to Extremes*, Oxford: Oxford University Press.

Ullmann-Margalit, E. (2017), *Normal Rationality*, Oxford: Oxford University Press.

Wood, G. (1998), *The Radicalism of the American Revolution*, Rev. ed. New York: Vintage Books.

Zuber, J. *et al.* (1992), 'Choice Shift and Group Polarization', *Journal of Personality and Social Psychcology*, **62**(1): 50–61.