Against Online Public Shaming: Ethical Problems with Mass Social Media

Guy Aitchison (Loughborough University) & Saladin Meckled-Garcia (University College London)

Social Theory and Practice, *forthcoming* (accepted 29 May 2020)

This is a pre-publication draft, please reference published version

Abstract

Online Public Shaming (OPS) is a form of norm enforcement that involves collectively imposing reputational costs on a person for having a certain kind of moral character. OPS actions aim to disqualify her from public discussion and certain normal human relations. We argue that this constitutes an informal collective punishment that it is presumptively wrong to impose (or seek to impose) on others. OPS functions as a form of ostracism that fails to show equal basic respect to its targets. Additionally, in seeking to mobilise unconstrained collective power with potentially serious punitive consequences, OPS is incompatible with due process values.

Keywords: Shame, shaming, social media, punishment, internet, respect

Consider this hypothetical case of shaming:

> A young scholar obtains a fellowship at a prominent university to carry out some innovative research on criminology statistics. A Twitter campaign builds up around claims by conservative activists for penal reform. They have found social media expressions by this scholar arguing that punishment is wrong, and in one case saying the reported murder of a police officer should be understood in context. The social media activists initiate an online campaign, with a petition, against her as a "cop-murder-apologist". She attempts to respond to different claims made on her Twitter timeline, but these simply repeat allegations that she is a cop-murder-apologist and their volume is so large that she cannot respond to even a significant fraction of them. Many of these tag important figures, newspapers, politicians, and potential employers. She is overwhelmed and shuts down her account. The campaign moves to writing to her new employer/university asking them to sack her, as her views display a poisonous disposition, especially to students who are children of officers. They highlight her work's failure to cite certain authors that have a different approach and methodology to hers. They investigate her family and discover she comes from a particular kind of anti-establishment background; her father having once been arrested after an altercation with the police. The campaign grows in reach, and some newspapers take up the story. Her university distances itself from her. Her fellowship is not, as would typically be the case, extended or renewed. Her name and 'cop-killer-apologist' become bound together in an enormous amount of online material. She finds participating in public discussions difficult given all her public interactions are tarnished with this label. She also finds

> getting further positions difficult, and some of her personal relationships are damaged as a result of her reputation. She has suicidal thoughts. She ends up working as a bus driver.

The above story is compiled using components from a number of real-life cases, reflecting common dynamics in the modern world of mass social media.[1] Real examples have involved activists of all persuasions, from right conservatives to left progressives, using social media to mount public shaming attacks on those they allege transgress moral norms that for these activists signal the transgressors fail basic requirements of moral personality for acceptance as an equal participating member in an important set of human relationships. In this paper we focus on this phenomenon in mass participatory social media, which we call "Online Public Shaming" (OPS), and we ask whether it is ever appropriate to engage in this special form of *reputational punishment*. We use the term punishment advisedly, as we shall argue that acts of OPS are directed at imposing a distinctive type of reputational cost on people and these constitute informal (non-state) and extrajudicial punishments that lack legitimacy. In socially-shaming acts, the punishment involves characterising people's personalities and moral characters as unworthy of participation in certain human relationships, and so as worthy of social exclusion. Imposing these punishments, we argue, attacks the victim's moral standing in a way that violates a basic form of respect we owe to all persons. In addition, the negative reputational build-up of digital media "pile-ons" can extend into disturbing "real world" social relationships such as those in employment and society. Our contention is that these actions are not simply exercises in freedom of expression with unfortunate consequences, or even desirable consequences where they are carefully targeted. They morally wrong their targets.

[1] The student Monica Foy was the target of an online vilification campaign for her tweets following the murder of an unarmed policeman; the philosopher Rebecca Tuvel was the subject of an open letter demanding the retraction of a journal article she wrote on "trans-racialism"; cancer expert Professor Tim Hunt admitted having suicidal thoughts following his shaming for allegedly sexist comments at a conference; literature professor Steven Salaita became a bus driver following the withdrawal of an employment offer at the University of Illinois due to his harsh tweets about Israel's actions in Gaza (Singal 2015; Singal 2017; LBC 2016; Salaita 2019).

We show that there is a cluster of recognisable moral wrongs that OPS instantiates, and which are sufficient to make perpetrators morally culpable. In this, we take up a distinct position in the literature. As we show, some commentators have defended online public shaming as having a valuable role to play in condemning the morally reprehensible, enforcing authoritative social norms (such as anti-racism) and drawing mass public attention to worthwhile political campaigns.[2] By contrast we argue that it is a moral wrong and a social ill when directed at individuals, regardless of its beneficial consequences. Our central argument is that the practice of OPS is an attempt to incite a public, collective punishment of people for the kind of person they are (their moral personality) and therefore mistreats them, stigmatising or dissuading their adoption of life goals, projects, and priorities that give shape to their lives. It subordinates their own development and pursuit of life priorities to collective judgements on their moral personality. The punishment of certain moral characters aims at their exclusion from certain social relations because of that character. It aims to punish by a public and collective social ostracism. Moreover, as we show, OPS is inherently incompatible with due process constraints.

Part 1 sets out the normatively salient features of social media and the phenomenon of OPS. Part 2 identifies key features of OPS as a distinctive practice in relation to other proximate categories of online behaviour, such as cyber-harassment, trolling and doxing, and offline behaviour such as malicious gossip and mocking. Part 3 pinpoints the moral wrong involved in OPS as a *form of punishment*, in terms of a violation of equal respect for persons and an incompatibility with due process. We also consider the responses that OPS might be merited by some acts and characters and that it might be acceptable if certain procedural conditions are fulfilled. Part 4 concludes with policy recommendations in the face of the ethical problems with the practice, including a "right to reply"

[2] While acknowledging its potentially destructive aspects, Paul Billingham and Tom Parr (2019) argue that online public shaming is permissible when enforcing moral norms subject to certain process-related constraints. We criticise their argument in Section 3. Jennifer Jacquet (2016) has offered a prominent defence of shaming, though her argument focuses almost entirely on the political campaigning value of targeting corporations and other organisations worried about bad PR, and so avoids the more troubling implications of shaming individuals which is our focus.

service for victims, recommendations for explicit clauses in social media codes of conduct and employment legislation to protect those who have been victims.

1. Special Features of Social Media

Digital mass social media are those media open (or relatively open) for people to join as participants, the communications on which are in principle available to the public and accessible by use of digital networks and devices. Users of such media create and exchange content, which includes written messages, but also pictures, video, audio and the like. They share information about their personal lives and ideas and come into contact with the lives and ideas of others. Social media platforms often grow and just as quickly decline in popularity depending on cultural trends and technological innovation. At the time of writing, well-established and prominent platforms include *Facebook*, *Twitter*, *Instagram*, *YouTube*, *Reddit*, *Snapchat* and *Weibo*. There are also messaging services, such as *Whatsapp*, and blog services, such as *tumblr* and *Medium*, Each has its own distinctive characteristics, but for our purposes there are some common features which pose the distinctive ethical questions we wish to address.

Social media allows for interactions, such as direct messaging, that have some features in common with private conversations as with posted letters, email or a conversation in one's front room. Other interactions have more in common with conventional forms of public speech at a meeting or rally, while others still resemble street interactions, including the online equivalents of heckling and mobbing. Such media platforms are private, in the sense that they are run as profit-making corporations with obligations to their share-holders, rather than being supplied as public amenities. Yet they also have many of the characteristics of a public forum. This is on account of their open, participatory character, offering an easily accessible space for communication among a large number of people who may be unknown to one another. Facebook, for example, claims to have over 2 billion

users, while Twitter claims 336 million regular users worldwide (Badash 2018; Fiegerman 2019). Here, views are articulated and criticised and one may be exposed to a diverse range of conflicting opinions. Because of the sheer numbers reached via these media and their interactive element, mass social media have become an important space for the shaping of public attitudes and the dissemination of ideas.

Mass social media can encompass a variety of social spheres, such as family relations, friends, colleagues, neighbours and people with whom we are linked by political or sporting allegiances, hobbies and interests. Online audiences may also include strangers with whom we share no obvious connection. As a consequence, postings on such platforms occupy an ambiguous position in relation to traditional conceptions of public and private. A Facebook post may be configured as "private" in the account settings, yet be shared with 1000 "friends", some of them unknown to the poster. Alternatively, a post on a platform such as Twitter may be shared to a small number of "followers", made up of friends and family, and yet still be publicly accessible to all. Even where users set their posts to fully private or closed to set followers, nothing prevents their messages being disseminated by others (e.g., as a "screenshot"). Given the platforms themselves are not formally public bodies, but private corporations, their own regulations and user codes of practice do not have to reach the same constitutional or human rights standards that governments do (Facebook 2020; Twitter 2020). This feature presents a challenge as to which accountability norms are appropriate. In addition, the impersonal nature of online interaction is known to have certain disinhibiting psychological effects, encouraging harsh and abusive behaviour that users would not contemplate in face-to-face interactions (Suler 2004: 321-326). This is especially so where there is the option of anonymity or adopting false identities.

A key point for our purposes is that social media is in principle open to mass public scrutiny and mass public participation and that it allows for a semi-permanent record of postings. It is this public

element, and especially its participatory character, that we shall focus on as it allows the mobilisation of a special type of informal power, giving rise to important questions of ethical appropriateness. This distinctively public character allows for what we shall call an *aggregative public effect*. This is where people use social media, expressing negative views about a specific person, to incite a special kind of public, collective repercussion. The speed with which information can be disseminated online, and the premium attached to being the first with a witty take or biting put-down, creates a rush to judge individuals before the context that may explain their actions can be established.[3] Force of numbers and the independent life that such claims may come to acquire over time can make responding difficult and costly for an individual financially, in terms of further disclosure, in effort, and reputation. The burdens of defending her character or simply explaining the situation can quickly become overwhelming. Crucially, online postings are accessible to the public long after any crowd action has taken place. Aggregative public effects typically generate a mass of public online material across a plethora of platforms, making it difficult for even the most determined victim to clear their name.[4]

The moral impugning of character in this context typically involves descriptions of a person as sullied and tainted, rather than stating facts or arguments concerning her views or behaviour. They are framed as someone beyond the pale, not to be trusted or engaged with. Characterisations such as "devious", "corrupt" or "dirty" perform this function, as do certain politically charged terms, such as "racist"; "anti-semitic"; "transphobic"; "bigot" or "traitor" (to their country) where the person is labelled rather than their actions and beliefs. These public characterisations of individuals using morally stigmatising

[3] Daniel J. Solove (2007) offered an early analysis of the threat Web 2.0 poses to the control people enjoy over their reputations, emphasising both the immediacy and permanency of online interactions compared to offline communication. We consider what this means for due process in Section 3.

[4] The EU has developed a "right to be forgotten" (*General Data Protection Regulation* (EU) 2016/679, Art. 17(2)), allowing the removal of past information about individuals from internet records, but this applies to links from search engines and for entities classified as data controllers. Social media platforms are different, as the free speech rights of those posting are in play, especially of journalistic outputs. The global coverage of the right is patchy, and its force in the US would be significantly shrunk by constitutional free speech protections.

labels for their character, whatever the basis for doing so, can place that individual in a difficult public situation when enough people join in. All of this can be below the threshold of anti-libel protections, given that these claims made by individuals might otherwise constitute fair comment or honest opinion (Rolph 2013: pp. 16-21). Importantly, our concern here is not with the veracity of the claims expressed in communications involved in OPS. They may be reasonable views. Our focus is on their use (Schauer 2015: p. 119). As we will argue, it is the intended purpose of such online comments that determines whether they wrong their target.

It is plausible that the mere threat of OPS can have a chilling effect on free speech. Faced with potential public, mass ostracism there is the danger that people self-censor and avoid discussing unpopular opinions, for the kinds of reasons famously identified by J.S. Mill. In *On Liberty*, Mill notes how public opinion itself can be coercive, imposing a crushing uniformity that deprives individuals of the liberty to experiment with new ideas and modes of life The "social tyranny" of prevailing opinion, he suggests, can be even more oppressive than punishment by public institutions, since "it leaves fewer means of escape, penetrating much more deeply into the details of life, and enslaving the soul itself" (Mill 2003: 76). In this way, minority positions can face suppression through collective online pressure, undermining collective deliberation over shared problems.

Our argument, however, does not focus on these possible (also wrongful) outcomes of OPS, but on the very treatment it inherently implies for its targets. The public actions involved in OPS, as we characterise it, seek to exclude and silence people because of a characterisation of who or what they are. They signal that a person is outside of the acceptable or decent moral community and turn her into an object of derision, rather than preserving her status as a subject and interlocutor. We argue that seeking to impose this cost on people is a wrong, independently of whatever direct harms it may inflict on the target, such as hurting their feelings, or indirect social harms to which it may foreseeably give rise, such as threatening their income from employment. To use a distinction familiar from the

theory of speech acts, here we are specifically concerned with the wrongfulness of OPS in terms of its "illocutionary" role — the type of action the speech act of shaming performs — rather than the "perlocutionary" effects it may have, such as the emotional distress of the victim and the degradation of public debate (Austin 1975).

Online public shaming has the following characteristics: 1) a person uses social media to publicly deride another person's moral character as having a feature that renders that character, in the shamer's view, as transgressive; 2) the shamer incites others to join or clearly frames her posting as open to them to do so, or she may join those already involved in doing this, adding to a cumulative effect; 3) the act is collective in the sense of people sharing an aim that requires others to join them to be effective, and that they act to bring about; 4) their characterisations of the target aim to publicly present their identity as being not worthy of participating in (certain) normal social relations, civil conversation, or debates as a moral equal; 5) the incitement to act in this way is pinned to a specific norm transgression by the target (though not necessarily illegal ones) which is treated as evidence of the morally faulty character that renders the target excludable. None of this need be legally defamatory, and the communication need not be with the target but can involve admonishments presented as communicating something to the target. Initial attacks can include, for instance, broadcasting footage of some behaviour deemed objectionable or retweeting a remark to one's followers with a call for repercussions. Where successful, such incitement rallies an angry cyber mob of other internet users, directing an escalating wave of derision, ridicule, or abuse at the target. Some of the communication here can escalate to calls for scrutiny of the target, her position, location, job, relations, personal life, and life of associates. Sometimes there is a call to exact retribution beyond social media derision, e.g., for the person's employer to be informed of her transgression with the aim of jeopardising her job.

There are two categories of agents who participate in an act of OPS: those who initiate or incite others and those who "pile on" by adding to the derision heaped on the target with more social media commentary or with likes and retweets. Agents who incite others to join in the shaming action are more culpable as they not only shame but also orchestrate the shaming, as with the person who responded to an American dentist hunting and killing a much loved lion, Cecil, by setting up the Facebook page "Shame Lion Killer Dr Walter Palmer and River Bluff Dental", which received 31,000 likes. Even if their attempt at shaming fails because no one else joins in, the intent involved makes the act wrongful. Among those who pile on, the degree of culpability ranges. There are, at one end, those who take more troubling active measures (such as joining in with the call to contact the dentist's workplace), those who amplify a shaming with retweets and comments, and those who take more passive actions (such as liking a page) and are therefore less culpable. The latter are nonetheless still guilty of participating in a shaming action, however imperceptible their contribution. As authors such as Derek Parfit have argued, even in cases of diffuse participation, moral responsibility can be allocated on the basis of the overall wrong that a group of agents are *together causing*, rather than on the isolated contribution that each individual makes (Parfit 1984: 79 -82). Where the participants share an aim that requires more than one contributing action (however diffuse) individual culpability is more plausible (Kutz 2000: 66-112). By joining with other social media users in this type of positioning of the target as a transgressor they participate in a collective act that, as we shall argue, is itself wrongful.

It is important to note that this act is *public*, both in the sense that it is open to public participation without presupposing any personal relationship with the target, and its aim is for society (in the sense of the public) to collectively impose the sanction on the target of exclusion from equal treatment in these relations: a collective public shunning, for having the moral character they do. Indeed, the collective shunning frames the target's personality as worthy of shunning from these relations.

In sum, OPS refers not merely to *what* is said, but *how* it is said. There must be some attempt to incite or join in an act of collective disqualification. The actions attack the public status of the transgressor as a member of the acceptable moral community, publicly depicting her as having a sufficiently morally deviant character that she is not worthy of interaction as an equal. The cumulative aims of the action are that the person is framed publicly as someone to be shunned and avoided, in such a way that precludes genuine conversation or discussion. This is a public sanction imposed on the target, and as we shall argue, constitutes a form of punishment.

2. Special features of OPS

The real-world examples of this are legion. In an early instance in 2005, a young woman in South Korea was filmed allowing her dog to defecate in a subway train. As a result, her personal identity was discovered and shared and a mass campaign of stigmatisation of her ensued (Solove 2007: 1 – 3). In a famous case, Justine Sacco, a PR executive, tweeted a clumsy joke about race to her 170 followers on a plane trip from New York to South Africa: "Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!". She later claimed her tweet was intended as a satire on white privilege. Upon landing, however, she discovered she had been fired from her job and was the number 1 trending topic worldwide. Given the shaming unfolded during her plane trip, Sacco's case came to symbolise the tendency of online collectives (and occasionally employers) to judge the character of those targeted without allowing an opportunity for a good faith exploration of actions. In some cases, shamers are themselves subject to reprisal attacks that are equal to or worse than the initial OPS. In one case, Adria Richards objected to what she perceived as a sexist joke about "dongles" and "forking" in a conversation between two male engineers at a tech conference. She photographed their faces and shared her story on Twitter as an example of sexism in the tech industry, leading to one of the engineers losing his job. This in turn led to Richards (a mixed race woman) being viciously

attacked online (including rape and death threats) and subsequently sacked after her employer was targeted with DDoS attacks.[5]

Consider, then, that OPS actions have some key features.

i) Shaming

OPS actions aim to publicly shame. Philosophers tend to identify shame as an emotion, albeit a social one in which the negative judgment of others interacts with one's own sense of self-worth and willingness to appear in public.[6] Within the literature on shame, it is common to draw a contrast with *guilt*, which is said to be focused on acts, rather than character. Guilt is seen as the more morally beneficial of the two, since it is possible to acknowledge a wrongful action, apologise, and move on or seek re-integration by making amends with the victim. With shame, however, one's very identity is supposedly at fault.[7] Many philosophers therefore hold shame to be a negative, destructive feeling that we would be better off dispeling from our inner lives. It is said by Martha Nussbaum to reflect the "narcissistic" and unattainable desire of the self for "completeness". Nussbaum links shame to rage and violence aimed at those blamed for one's deficiency. Other philosophers claim, by contrast, that shame has a valuable role to play in moral life. In recent debates, Julien A. Deonna, Raffaele Rodogno and Fabrice Teroni have argued that the desire to avoid shame experiences can play a worthwhile instrumental role in motivating self-improvement (Deonna, Rodogno & Teroni 2011). Krista K. Thomason, meanwhile, proposes that a disposition to feel shame reflects an appropriate sensitivity to the perspectives of others and is thus constitutive of a moral practice of mutual accountability among moral equals (Thomason 2018).

[5] These and other examples are documented in Ronson 2015.
[6] Some claim shame requires a (real or imagined) audience, e.g. Taylor 1985. Others suggest that it can be experienced as a private emotion, e.g. Thomason 2018.
[7] Nussbaum notes how, in contrast to shame, guilt can be "potentially creative". She connects it with reparation and forgiveness, given its focus on the wrong or harm of individual acts, 2004, pp. 207 - 209.

For our purposes, it is not the emotion but a related practice of *shaming* that matters. The presence of shame as an emotion is neither necessary nor sufficient for an activity to count as shaming. What matters, for our argument, is the manner in which some agents characterise a certain element of moral character as shameful, meaning that agents who have this trait are not worthy of the self-respect necessary to take part in public discourse. Whilst shame as a feeling may ensue from such actions, it is the characterisation of a person as shameful, unworthy of public participation, and beyond the pale, that is the core of this practice. Whatever her emotions, publicly *shamed* individuals suffer a loss of control over their public identity due to others' perception of their character as fundamentally defective, rather than opprobrium being directed at their actions, which are within their control.[8]

Public shaming is a communicative social practice in which people are framed publicly as being outside of the community of morally acceptable persons, however they feel about it. The public shaving of the heads of women, believed to be collaborators in the Second World War, would be a paradigm case of public shaming (Duchen 2000). The shaved head represented a publicly accessible message that this person is stigmatised and excluded from the normal moral community and (whatever the individuals felt about it) they have this identity publicly assigned to them. In contrast to offline shamings, OPS involves what we have termed an aggregative public effect. This increases the risk of a dangerous escalation and heightens the exclusionary power of a shaming, which is open to immediate mass participation by a global audience unknown to the target and unaccountable to them. OPS affords even less opportunity to escape from public gaze than a shaved head. While the shaven collaborator suffers humiliation, her hair will eventually grow back. With OPS, the target is tagged with a character label, such as "bigot" ("traitor", "racist", "cop-murder-apologist", "anti-

[8] The anxiety associated with shame as an emotion, and the difficulty shamed people have with re-integration, may stem from this loss of control over how one appears before others. The degraded part of the shamed person's identity is taken by others to define their whole being, conflicting in a fundamental way with how they understand themselves. For accounts of shame that emphasise this loss-of-control aspect, see Thomason (2018) and Velleman (2001).

Semite", "transphobe", etc.), and publicly associated with that label in a long-lasting way. A public record of the target's misdeeds is stored and made accessible via search engines, along with the insults and speculative commentary of countless others. This public record is semi-permanent, individuated (being tied to a specific name and identity) and prominent, especially where a shaming attracts the attention of high-profile media. An internet search of "Adria Richards", for instance, returns 27,300 results, with the first page returns of Google displaying articles in high-ranking outlets, such as Wired and Forbes.com. There are no clear bounds set by the shaming action in terms of who can participate or whether it is restricted to online interactions only. Control over its expansion is not incorporated into the medium or the act itself.

This participatory aspect connects with an important characteristic of mass social media. Participants can post without redress (so long as the posts stay within the law and avoid clear defamation). In contrast to best journalistic practice, targets do not have a right to put their side as part of the post, nor is there a right to redress for postings (Frost 2015: 113). Nor do they have a right, or (in most cases) ability, to respond in a way that has an equal and proportionate force to the numbers involved in stigmatising them. A "conversation" of tens, hundreds, or thousands can take place about a person, in which she is labelled with a shaming term and which excludes her from participating in any meaningful way. This effect is produced by the sheer numbers that can become involved but also by the act of labelling her in OPS communications that are not framed so as to open discussion, acting only as condemnatory labelling. The label, and those using it, signal that the target is "cast out" or not a participant with some say in what is happening to her. Sometimes the explicit term used is that she is "cancelled".

ii) Norm-patrolling

OPS actions are presented as stigmatising individuals because their views or actions transgress certain social norms that the shamers uphold as moral red lines. This may be done as a lesson to others or as confirmation among the shaming group of what is acceptable. It may even be used to try to extend the reach of such norms.[9] The key aspect for us, however, is that the activity is stigmatised through a form of social reputational cost, because of the perceived norm breach. The young woman in South Korea, for example, breached certain norms governing acceptable behaviour for when your dog defecates in public space.

It is this patrolling of norms by stigmatising a person that distinguishes OPS from other online activities with which it sometimes shares features, such as online "trolling" where internet users engage in provocative and digressionary interventions in online discussions with the aim of disruption or simply obtaining a reaction.[10] It also differs from harassment, stalking, and "doxing" (involving breaches of privacy where personal information about a person is released online) and online hate speech. Cyber-bullying (or -harassment) involves persistent online aggression that is calculated to inflict emotional distress (Citron 2014). It may involve abuse directed at an individual, using doctored images or videos of them, contacting their friends or family, and at the extremes, rape threats, death threats or death wishes. As with OPS, cyber-bullying is often an aggregative, mob-based activity. Indeed, OPS often gives rise to cyber-bullying down the line where internet users heap further torment on the target with persistent acts of aggression, all in the name of enforcing a social norm. Where this is done publicly, an action may qualify as *both* cyber-bullying and OPS. Alternatively, a mass shaming can provide cover for bullies — who may have no interest in the moral issue involved —

[9] Norm theory identifies norms as informal standards among a group that are a matter of common (often tacit) knowledge. Some scholars propose that harmful social norms can be modified or removed through expectation-change by debate of the need for change plus public declarations by relevant (influential) parties (see e.g., Bicchieri 2016: 219 ff & 156 ff). Other accounts of norm change concur on norm changes happening when one gets enough people to accept a new norm or new interpretation of a norm, Brennan, Eriksson, Goodin & Southwood, (2013): 94 ff., and this may be achieved through the threat of informal sanctions (pp. 97 ff).

[10] The legal scholar Kate Klonick also sees shaming as connected to norm enforcement (2015): 1029; by contrast, Laidlaw uses a wider definition which involves shaming where no norm-enforcement is involved including harassment, doxing and other forms of unwanted exposure, (2017): 3.

to inflict emotional distress on the target for its own sake.[11] However, unlike OPS, it is not inherent to cyber-bullying, or indeed cyber-harassment, trolling, or hate speech, that these actions are responses to, or even explained by, norm breaches. Clearly, drawing precise boundaries around what counts as cyber-bullying is difficult in practice and the activity raises its own moral and legal questions, touching on the appropriate boundaries of legitimate speech. Yet its objectionable character is relatively uncontroversial, and in fact addressed in codes of conduct and legal prohibitions.[12] There is less consensus when it comes to OPS.

Furthermore, not all cases of online shaming qualify as OPS under our conception. A simple act of disclosing personal information can lead to feelings of shame (or be intended to lead to those feelings), as with the publication of intimate information or images, such as naked photos. But such actions are not always connected to a norm breach since personal attacks may aim to harm and humiliate a person, rather than to label them as shameful. There may of course be a mixture of intentions in any one action. Cases of "revenge porn" can be ambiguous for instance, involving a mixture of both personal humiliation and OPS. When someone non-consensually posts sexual material of a former romantic partner online, which had initially been shared in confidence, this may be done to exact personal reprisal or perhaps for financial motives, rather than to exact a reputational cost for transgressive behaviour. The much greater prevalence of women targeted by revenge porn however (92% of victims according to some research) suggests that the phenomenon often involves at least implicit appeal to traditional norms of female modesty, with those victimised being taken to have "deserved" their punishment for having behaved in a sexual way.[13]

[11] As one man who was jailed for cyber-bullying in the UK put it, "The irony of it all is that I wasn't even passionate about the subject or the people I was bullying. I was simply bored, saw what was trending, and leaped on to the bandwagon." (Smith 2015).

[12] See for example Citron (2014) and the collection of essays in Levmore and Nussbaum, eds. (2010).

[13] Another contributing factor to revenge porn is that women are more likely to be objectified in general. See Uhl (2018): 50-68; and Citron and Franks (2014).

Finally, as we have mentioned, OPS is performative: in publicly and collectively characterising a target as an "anti-Semite", "cop-murder-apologist", or a "traitor", as part of an online mass action, shamers not only engage in discussion but also perform (or intend to perform) acts of shaming and exclusion. They increase the numbers of people using the label to characterise the individual and her character as lying outside the group of people with whom it is morally acceptable to converse, or those who are accepted as equal participants in public interactions. Characterising someone as not worthy of membership in moral society can also mean that the shamed person is open to a number of other actions, as a legitimate target. These can add to the stigma, as with the use of abuse, or add to the sense of being cast out of moral society, as where attempts ensue to affect the person's social relations such as their employment or membership of associations. The use of insults, such as "idiot", "loser" or "bastard", generally has the aim of hurting, offending, or humiliating another (Archard 2014). It is not an inherent part of OPS but can ensue or accompany it because the victim is not considered protected by the rules that apply to those respected as equal members of moral society.

iii) Imposing a public and collective cost

It is inherent to our characterisation of OPS that it involves stigmatising a target in a way that treats them as not being a legitimate member of a community of equal participants in a subset of normal human relationships that are important to their well-being and sense of self-respect. They are "cast out" in the sense of being identified as someone to exclude from normal relationships. This can be restricted to exclusion from discussions and exchange of ideas but can also be extended to employment relations or even being served at retail outlets.[14] This feature of OPS makes it a special

[14] There were calls for a man not to be served by any business outlet because of his use of racist abuse on a plane (Hovellin' Hermit 2018).

case of attempted *ostracism.* In Ancient Athens, ostracism was a formal civic procedure used to coercively expel unpopular citizens from the physical territory of the city-state (Forsdyke 2005: 144 ff). OPS is not performed officially by the citizen body, but by an informally aligned group of individuals, whose actions aggregate to collectively shun a person through applying labels to her associated with having a morally unacceptable character. The target is not only denounced, but in the process their contributions or responses to the action are treated as morally immaterial by the shamers. The action can even involve an invitation to employers, schools, universities, political parties and other civic organisations to performatively cut ties with or denounce the shamed person.

Where successful, OPS is the exercise of informal social power, in that collectives of internet users are mobilised to impose them. It is informal in that it depends on voluntary actions by those participating and is exercised extrajudicially and independently of the formal coercive powers of the state. It is also not guaranteed to build up to a mass action, but where successful it does. It is a publicly exercised power in that it employs a medium that allows participants to join regardless of their relationship to shamers and shamed, indeed incites people to join in regardless of their relationship or lack of it, to the target, and creates a record that can be relatively freely accessed by (or reported to) the general public. Nothing is in place that restricts the action to personal or private interactions by individuals. It is not even necessary that the target be digitally "present" on the social media platform or even know about the shaming for it to occur. Thus a person cannot avoid OPS in the way they can avoid a private interaction or relationship, and more importantly they cannot avoid its public character. Whilst it is informal, the use of social power in OPS shares some key traits in common with state-backed punishment.

As with state punishments, OPS actions purport to speak in the name of the moral community.[15] In line with standards conceptual definitions of punishment, it aims to impose a loss (pain, harm, removal of freedoms, removal of rights), by those with (claimed) authority to do so, on those who have (allegedly) breached a norm, and where the deprivation does not merely lead to a forfeit but is accompanied by censure or condemnation.[16] The intended exclusion in OPS is a significant loss of status: as an equal member of the moral community, and it is not intended to be limited to a particular space or group, as would the loss of status within a private club, say. It has the potential to negatively impact self-esteem, health, personal relationships, institutional relationships, and employment. By implication, those inciting it or piling on claim the moral authority to do so, and the status loss is inherently censorious and stigmatising. In this way, OPS is a genuine extrajudicial attempt to punish individuals.

It is also worth noting that being subject to OPS is not prevented by social advantage. The impersonal dynamics of social media entail that shamers are prone to under-estimating both the power they have over their target and the harm inflicted by their actions (Norlock 2017). The harmful effects of OPS might be felt more acutely depending on the resources and social standing of the target. Women and ethnic minorities may be singled out for particularly vicious online attacks, as with the counter-shaming of Adria Richards.[17] Those at the lower end of the social and economic hierarchy also tend to be more dependent on their communities for material and social support and are therefore especially vulnerable to being excluded (Massaro 1997: 645). Wealthy individuals, by contrast, can

[15] Feinberg (1965): 397-423. Note that our account is consistent with any theory of punishment and its justification, so long as the theory accepts that a certain kind of socially-imposed cost that stigmatizes the costly action is a punishment. Our critique in the below is consistent with those theories of punishment (e.g., the communicative theory) that see a problem with the mere instrumentalisation of people through a punishment regime, viz. Duff, (2001), eg., Dagger (2011).

[16] The definition of punishment (rather than a particular theory) in the literature is consistent with OPS being an instance of it. See Flew (1969); Duff, (2001), pp. xiv-xv; Duff points out that there is nothing essential to punishment that means it must be performed by the state, p. xiv, ibid; See also Duff, (1991), pp. 151-3; others have talked about the informal, and potentially positive, application of "rough justice", Goodin, (2019).

[17] Ronson notes that the most violent attacks are often reserved for women and minorities, while wealthy white men are the most likely to recover after a shaming (2015).

call upon the services of dedicated "reputation cleaners" to erase any trace of online controversy (Wood 2013). Nevertheless, this does not make them immune from successful shaming acts.

3. The Wrongfulness of OPS

An initial reaction to our discussion so far might be to say that where online public shaming is inaccurately applied – involving a misjudgement about what was actually said or a misrepresentation of the speaker – it is unacceptable.[18] Where the judgements are correct, it is not. Perhaps, as with defamation, only where claims are false does the reputational cost amount to a wrongful loss, otherwise the punishment fits the crime. Pressing on OPS as a type of punishment, this objection might say that even extrajudicial punishments are sometimes appropriate as a response to wrongful action.[19] They might also be justified as "justice-forcing", pushing for the bringing about of just laws where formal justice is lacking (Goodin 2019, pp.88 ff). Some authors defend shaming as a legitimate penal measure (Goldman 2015: 415), and others argue that the point of shaming online is not inherently tainted, only that important safeguards are necessary (Billingham and Parr 2019). In what follows, we argue that the cluster of features inherent to OPS as a practice make it morally unacceptable. Our key claims are that OPS is a failure of fundamental respect for people as separate human beings with distinct lives to live and inherently incompatible with due process.

3.i. Respect

Consider Stephen Darwall's notion of "recognition respect". This reflects the equal moral standing of individuals as persons (Darwall 1977). Recognition respect requires that we treat people in line with this standing. A key component of that respect is respecting people's right to live lives according

[18] Many of the examples in Ronson (2015) are at least in part due to misrepresentation or misunderstanding. This also appears to have been the case with Tim Hunt's case (Foreman 2015).
[19] See also Simmons 1995: 221 ff.

to their own lights and ends, so long as they do not have the end or project of harming others' rights. For this reason, a significant range of liberal political theories integrate respect for people as separate individuals with distinct and separate lives to live (albeit with some social responsibilities that makes this possible for all). Social arrangements, on that kind of view, ought to be neutral with regard to people exercising a capacity to choose their life values, goals, and projects, and how to prioritise these at different points in their lives. No one is required to have any specific kind of character in such arrangements, except to the extent that someone's character needs to be compatible with respecting the rights of others in one's actions. Indeed, within the range of actions compatible with respecting the rights of others, it has been argued that it is even permissible to do moral wrong and by implication to be disposed to doing such moral wrong. This is because of the overriding weight given to respect for others to determine their own life priorities and values (Waldron 1981). Importantly, people can hold views that perhaps a large section of society holds to be deeply erroneous, or even morally base. Thus, liberal societies can accommodate people who believe homosexual relationships are wrong, so long as these people do not breach public rules and obligations. It is this kind of basic respect for people that underpins liberal neutrality in the framing of institutions (Larmore 1989: 580-581; Meckled-Garcia 2017).

It is true that liberal authors, like Mill, accepted it as proper that citizens should concern themselves with the development and well-being of others in society. We may, for Mill, judge others' bad behaviour, even holding them in contempt, refusing to socialise with him, encouraging others to do the same, and withholding certain opportunities. But these responses are understood as the spontaneous, uncoordinated outcome of individual behaviour, not something intended to be deliberately and collectively inflicted as a form of public punishment (Mill 2003: 140 - 141).

We can now see what is objectionable about OPS. It targets individuals to impose a cost on them precisely based on a characterisation of their moral personality, choices, and character. The cost in

question attaches not simply to an action, but via an action to her for being a type of person. Indeed, one cannot disentangle the idea of being publicly shamed like this from a stigmatisation of character and ostracism for it. Even where OPS appears directed at specific actions, its mode of punishment (ostracism) still punishes character, as unacceptable in a moral community. If others cease treating a person as a moral equal worthy of participation and indeed characterise her as such, whatever the trigger, this is an indictment of her worth as a person. We should emphasise that this is all within the bounds of respecting others' basic rights. Clearly, criminal activity or basic rights violating actions (such as incitement to murder) can legitimately lead to calls for action and refusal to communicate with the person. But here, the point of the action is not enforcement by ostracism, but in isolating a criminal or forcing the state to act justly.

We should emphasise here the distinction between the practice of shaming a person and an act that might lead her to feel shame. When someone says to a person : "You ought to be ashamed of your sexist behaviour!" that is a criticism of the action and expresses the hope the person will come to feel a certain reaction to it. "X is a sexist; pile on and cancel them!", however, is public and character-based. OPS is categorically distinct from even the strongest or most emotive criticism of someone's actions.

In a co-authored paper, Paul Billingham and Tom Parr have argued that shaming people online can be justified as a means to enforce "morally authoritative" social norms in cases where legal regulation would be inappropriate. In their analysis, shaming is an "informal sanction" that fortifies the internal motivations individuals have to follow social rules (Billingham and Parr 2019: 5). It accomplishes valuable tasks, they suggest, censuring wrongdoers, deterring others from transgressions and reaffirming public support for worthwhile standards of conduct, such as antisexism. In their analysis, however, for online public shaming to be justifiable it must meet certain constraints, such as being

proportionate to the offence committed, while allowing for the accountability of shamers (through curbs on anonymity) and the social reintegration of the target.

There are several problems with this analysis. First, the authors specifically confine their argument to norms that are "morally authoritative", omitting the important matter of who gets to decide which norms are sufficiently authoritative to warrant that people should be collectively punished for their violation and how such decisions get made (Billingham and Parr 2019: 3). With no formal accountability procedures, the danger is that only behaviour which the most powerful social media groupings object to and which happens to catch their attention will be punished. Even where there is broad agreement on normative questions at the level of abstract principle — such as norms against racist speech — there is often empirical disagreement about which cases these principles apply to (as the authors themselves acknowledge). In the case of Justine Sacco, for instance, some saw her tweet as evidence of racism and others as a satire on racist attitudes. The problem of arbitrariness inherent in OPS cannot be bypassed simply by stipulating that it be used for "morally authoritative" social norms.

Second, Billingham and Parr discuss online shaming as a "burden"; a type of informal punishment which individuals are morally liable for as a result of bad behaviour (2019: 8, 11). However, there is little sense of what precisely makes shaming so burdensome for individuals or the troubling power dynamics involved. Their focus is on actions (not character), suggesting an interest with guilt, rather than shaming as such with its associations of collective denigration and moral exclusion. For Billingham and Parr, the personal stigmatisation aspect of public shaming is incidental to the practice rather than being a core component, which is what makes their "proportionality" criterion seem plausible. Indeed, it seems sufficient for them for something to count as "shaming" if it involves public criticism of a person's actions online. If this is all they have in mind, however, the account is highly inclusive. It thus risks subjecting harmless online speech to a wholly excessive set of

interactive constraints. At the same time it ignores the specifics of OPS that make it especially objectionable as a form of extrajudicial punishment. Of course, one could hold that if OPS enforces a valuable social norm, then the reputational damage it involves is always proportionate. But that would imply a particularly draconian and anti-liberal vision of social coexistence. If the point of human society involves mutual respect for people to determine their own ends, values, and priorities (even where these are mistaken), punishing moral character (as opposed to punishing rights-breaching actions) is morally wrong.

The value of mutual respect also speaks against OPS being justified or permitted when the target is from a privileged group (the "punching up" argument). If OPS breaches basic respect, then it does so whether punching up or down. As discussed, some vulnerable groups may be more vulnerable to the effects of OPS, and some less vulnerable people may have resources to shield against the consequences. But that does not mean that those less vulnerable or more privileged are more deserving of stigmatisation. They, as people, are entitled to basic respect for their personality, regardless of the position they hold.

Critical activities, such as attempts to convince or remonstrate with a person, are consistent with this kind of respect so long as they are not coercive. They treat her as a moral agent capable of reflecting on her behaviour and reforming it, but also as entitled to make decisions about her values and prioritise her ends. Even inviting someone to feel shame for an act is consistent with this kind of basic respect. By inviting a person to feel shame (rather than shaming them), we ask them to voluntarily reflect on whether their habitual behaviour matches up to important moral ideals and the type of person they aspire to be.[20] Consistent with this baseline of recognition respect we may even have a very low opinion of a person, denying them "appraisal" respect and even holding them in

[20] See the discussion of "constructive shaming" by Nussbaum. To open the door to genuine moral improvement, she proposes that invitations to feel shame should appeal to ideals from a shared political culture while being non-insulting, non-humiliating and non-coercive (2004), pp. 211 - 216.

disdain (Darwall 1977). It is not consistent with respect to call on others to join in excluding someone from social relations as a "bigot!" or publicly excluding her from participating in certain communal relationships — indicating she has no right to be there — because of her moral character. Note that none of this precludes withdrawing from discussion (choosing one's friends and interlocutors) with those one finds objectionable. It is incitement to do this collectively, using social branding, that matters here.

Another author, Thomason, agrees with our claim that invitations to feel shame are less objectionable than shaming because they lack its public aspect. Yet how she understands such interventions differs in important ways. We have in mind here forceful criticism designed to urge a person into serious self-reflection, without necessarily aiming to impose shame on them. "Do you really want to be known for this pathetic behaviour?" or "You're a disgrace to your profession!" would qualify. By contrast, Thomason regards invitations to feel shame as a more thoroughgoing provocation, including expressions of disgust and ridicule, where the intention is to catalyse shame in the target. It follows that, for Thomason, invitations to shame should be reserved for a specific set of cases, most notably those of "moral self-defence" against arrogant individuals whose belief in their own superiority means they would dismiss any legitimate criticism directed their way (Thomason 2018: 187 - 190). It seems reasonable to believe, along with Thomason, that uncivil interventions, such as ridicule, may have a legitimate role to play in some online interactions, so long as they are not about degrading or excluding someone. Nonetheless, it would seem more appropriate to characterise these acts as *provocations to shame*, rather than invitations, since shame is being imposed in a way that bypasses the voluntary cooperation of the target.

These points also help to distinguish OPS from other off- and online activities that attack persons and their personalities, such as public mocking and malicious gossip. Those activities are not necessarily aimed at exacting a public punishment, using an inherently public medium, based on personality.

Note that we do not say OPS is the only form character-based ostracism can take. Indeed, some forms of mockery and gossip intend social ostracism. However, mass social media is particularly suited to this activity and to public participation in it, while the speed at which OPS takes place and its one-sided character (pitching individuals against amorphous collectives) makes it much tougher to withstand and challenge.

3.ii. Due Process

As we have argued, the form of ostracism involved in OPS, with the intended social stigma, can be seen as a type of informal punishment. Formal penal systems that respect due process values will include at least the following key due process features: i) that the penalties applied are explicit and transparently applied, and that these have been arrived at through a social-deliberative process that makes them the genuine penalties of a political community, in which institutions are accountable to the community for their penal standards; ii) that these penal rules respect fundamental human rights; iii) that they are proportionate (in some rationally definable and defensible sense of proportionate); iv) that a trial or tribunal system for applying the penalties exists and that this gives those facing a potential penalty an opportunity to participate in the decisions being made, including by defending themselves against the charges or accepting them. In this way, due process norms for the application of penalties exist to guarantee *a fair balance* between the interests of individuals (who face the penalties of a justice system) and the social goals of that system.[21] Seeking that balance is a way of showing respect to individuals even if punishment goes against key interests they may have. Typical vigilante "justice" applies punishments that fail these tests. However, we are agnostic in this paper as to whether some version of informal justice could preserve these features. What matters is that this

[21] See for example, Council of Europe, "Guide on Article 6 of the European Convention on Human Rights", 2019: 27 ff. We are here using 'penalties' in a broad sense to include rehabilitative measures. See also, Duff, (1991):. 110 ff.

requirement to seek a fair balance applies to any persons acting collectively in applying public (social) punishments. Our argument is that OPS actions inherently fail (i), (iii), and (iv), and so cannot, even if it were a formal punishment, secure a fair balance between individual interests and social ends in applying punishments.

The norms that get enforced through OPS, are not subject to social approval, transparency, or accountability processes, nor is the form and nature of their enforcement. In his discussion of how societies come to embrace certain doctrines as the unquestionable truth, Mill noted how we have no assurance that society will coercively enforce the correct values, emphasising the pure contingency of historical affairs. Moral questions, Mill noted, can be decided by popular prejudice, the tastes of the ascendant social class, or the outcome of battles between rival national powers (Mill 2003: 77 - 78). Similarly, the decision on which opinions go unpunished online will be in the hands of whoever happens to hold sway over the most powerful or relentless Twitter groupings. At best, what gets to count as a norm, the transgression of which calls for shaming, is sensitive only to its ability to gather enough people and social media reach to successfully impose a label on an individual and thus impose a reputational cost on them. The idea that OPS might be associated with enforcing useful social norms sidesteps the important question of the decision procedures and the legitimacy of the process leading to such norms being adopted for the purpose of punishment.[22] In the use of public penal measures, legitimacy comes from established collective political decision-making processes that are transparent and clear enough to act as (or aspire to be) fair public norms, not just popular norms. The only exception to the need for this legitimacy condition might be extrajudicial actions to prevent basic rights violations. Where that is not the case (and by definition it is not the case for OPS), it does not matter that the norms being enforced are seen as morally valid. To be legitimate grounds for enforcement, they would need to be public norms, in ways that make them transparent and allow

[22] While our argument focuses on the arbitrariness of which norms get enforced, Thomason highlights how shaming is suggestive of moral arrogance among shamers who grant themselves illicit power over others. Shamers position themselves "moral police", overlooking their own inevitable moral flaws (2018: 198).

accountability for their application and to show basic respect to the target as an agent, and not merely an object to be acted on, by allowing her participation in the decision-procedures relating to the application of a punishment to her (e.g., through an opportunity for her or her proxies to make representations to those making the decision or to state a defence). It is also worth noting that whilst the (perceived) transgression of a norm may trigger OPS, the punishment is for having a certain moral character. Enforcing that kind of norm is incompatible with legitimacy for substantive, rather than procedural, reasons as set out above.

Similarly, OPS has no filter of proportionality for the reputational costs involved. Information can be globally disseminated instantaneously, in the amount of time it takes to send a tweet, so that a specific incident or remark judged to transgress a norm has the potential to escalate into a mass digital pile-on in a matter of minutes and to remain in the online record for a long time, irrespective of how minor this transgression might be. It is this unbounded (uncontrollable) nature of OPS together with its permanency that ensures the reputational costs of OPS are invariably disproportionate. Prior to the internet and mass social media, any information pertinent to an individual's character flaws tended to be "scattered, forgettable, and localized", rather than being available for global dissemination online in a way that "is permanent and searchable", and not inherently confined to specific boundaries, as Daniel Solove has noted (Solove 2007: 4). A real-world space, such as a subway carriage, was still public, but it carried with it a certain degree of anonymity (Solove 2010). A person guilty of a norm breach in such a space would not necessarily have had their identity exposed and, in the event that they did, their actions, and condemnation of them would not be publicly available in a potentially unbounded way across time and geography. If someone's bad behaviour was the object of social commentary, they could still hope to move on and develop as a person. Today, by contrast, someone whose picture is posted online is readily identifiable and hence they face being rebranded as socially deviant in a semi-permanent, publicly accessible record, without inherent boundaries.

For OPS to take place, there is no institutionalised requirement that those piling on are in any way involved with the initial incident or, indeed, that they have any relationship whatsoever with the person being targeted. It costs little to voice opinions on social networks, with no barrier to participation according to relevant knowledge, capability, or familiarity with the facts or the person, let alone guidelines for imposing penalties.[23] Social media can produce difficult-to-control, unbounded, information cascades, with people relying on hearsay and assertions by others as grounds for their own beliefs about, and actions towards, a person (Sunstein 2010: 91–106). The reach of social media, with the use of hashtags and groups, make it easier to rally and aggregate online mobs than real life mobs. Established broadcast and print media will often compound the situation by using the fact that a shaming has "gone viral" as a reason to report on it, contributing to the cycle of shaming and media commentary. Moreover, those inciting OPS cannot control what "goes viral" or its wider effects. This is its unbounded nature. All this challenges the predictability, proportionality, and therefore fairness of the imposed costs.

In the context of the modern criminal justice system, legal scholars have argued that shaming punishments fail proportionality tests, which require that the severity of any punishments should reflect the seriousness of the offence. Shaming by the state, it has been argued, relies on the volatile and uncontrollable urges of the crowd for its effects and it is therefore not possible to "calibrate" its severity (Whitman 1997: 1055). With no definite end-point to a shaming (what we have called its unbounded nature in the case of OPS) the effects can spill over into multiple different areas of an individual's life, including the shaming of innocent friends and family. Frequently, anyone with ties to the shamed person becomes tainted by association. In the context of social media, the proportionality problem that is inherent to shaming is exponentially amplified.

[23] This also creates opportunities for false and malicious accusations. For example, a man used a website that was set up to publicly shame people for their racist views to frame his ex-girlfriend, which led to pressure on her employer to fire her (McDonald 2014).

The idea of OPS cannot be separated from those aspects of it that conflict with due process, and much less so with basic respect. The conflict is inherent to ostracising public punishments for being a certain kind of person, and to such shaming punishments being exacted through mass social media. For these reasons OPS is an ethical wrong, and an unreformable one at that.

Caveats

None of what we have said rules out free and open criticism, even strident and indignant criticism of others on social media, although offensiveness and abuse might be ruled out by other valid inter-personal standards. The phenomenon we are criticising is that of imposing a reputational cost, inciting or participating in using informal collective power to do this, to stigmatise a person as excluded because of judgements about her moral character. Criticising her or her actions is only unethical on this standard if it is part of or a proxy for that kind of stigmatising. This kind of excluding behaviour that is inherent to OPS — the online analogue to ostracism — should also be distinguished from other forms of exclusion that are in fact legitimate. No one is obliged to engage in conversation with, associate with, or befriend anyone else. So, if people declare a wish not to participate further in a discussion, mute, or block an interlocutor, that is not an act of OPS. It is only inciting or joining informal social stigmatising of that person's moral character, and thus imposing a reputation cost, that counts as OPS, and is wrongful.

OPS can also be distinguished from online actions aimed at warning others about a person's behaviour where there is no better means to issue such a warning and where the dangers are real and significant. The #MeToo movement, for example, highlighted how pervasive the problem of sexual harassment — especially by powerful men — can be within specific industries. While much of the social media conversation was about raising awareness, some postings identified specific named individuals for predatory behaviour that was abusive and threatening (Khomami 2017). Many such cases would

count as fair warning to others, rather than objectionable attempts at shaming. Communicating illegality with the aim of bringing the suspected criminals to justice is also not OPS, given its aim. Consider the exposure of genuine concerns about the moral character of those seeking positions of trust, power, or authority, where that character is a qualification or where it will make a difference to how the role is executed, such as those seeking political office. This is not OPS because it is about qualification rather than punishment. Finally, even deprecatory characterisations of personality are not in themselves acts of OPS. Saying to someone "You, sir, are an anti-Semite!" in an online argument might simply be an exclamation, rather than incitement to OPS. Line-drawing judgements will of course be highly context sensitive, which is not to say there is no line to be drawn.

Two objections

An obvious objection to our view is that some people deserve to have their characters stigmatised and to be shunned as a fellow communicator because of the harms they have done or the extreme nature of their views. In fact, the response goes, this is a valuable tool in the struggle against toxic views and toxic people online. Someone holding racist views or a homophobic person, for example, should be despised, as should the Nazi sympathiser, or the extremist propagandist. Stating that a racist is a racist is just a statement of (believed) truth, and we cannot ask people to suppress the truth. In an often-used trope online, "free speech does not mean being free of the consequences of one's speech", and one such consequence is being called out as such, albeit by very large numbers of people.

It is worth noting that liberal concepts of justice and human rights have classically always allowed the prosecution of people because of acts of violence, coercion, or incitement (which includes inciting hostility towards specific groups in the form of hate speech). But where that is not the case, liberal views have not called for the punishment and repression of people for their views, even where those views are low quality, offensive, or even demeaning to others. To a large extent (with clear

imperfections in the application of such principles) liberal societies have sought to uphold this distinction. If our argument is right that OPS is a kind of informal punishment — a real cost imposed by informal collective action — then it is supremely anti-liberal. It punishes people for the content of their views, and more so it punishes them for the content of their tainted character, as flagged by their views.

Another model of the social media world, however, sees it not as a political community but as *modus vivendi*, or even a Hobbesian state of nature (and so state of war). It might be argued, then, that what matters is defeating reactionary views and bolstering progressive ones by whatever means necessary. The use of OPS is then just another tool in that war and progressives need all tools at their disposal. However, the baseline for participating in what is supposed to be a public conversation should be some baseline of respect for others as equals. This form of recognition respect is owed to individuals as persons irrespective of our opinion of them. Seeing people as instruments towards achieving progressive goals, whatever the cost, fails in that basic standard of respect.

4. Implications of Principles for Policy

Producing and posting lists of people that declare them to be morally tarnished in some way, online letters collecting signatures condemning a private person as morally tarnished, using words that incite people to join a character condemnation ("….bigot"; "left wing fascist") and to take further actions ("get them fired!"), are paradigm examples of OPS. They are to that extent wrongful. Take the website "Rate my racist professor", which encourages students to submit anonymous ratings of professors at North American campuses for purported racism across such categories as "Anti-American", "Anti-Israel" and "Anti-Immigrant". The professional-looking interface compiles these ratings into a "Racist score", attached to a personal profile, with the worst purported offenders featured in a "Racist

hall of fame". The site claims authority for these characterisations on the basis that submissions come from "verified academic e-mails" with (thin and contestable) supporting evidence provided in the form of quotations from "third party sources", including commentary on social media sites. Those featured are not invited to respond or to contextualise their statements. The clear intent is to publicly degrade their reputations, while encouraging students and colleagues to cut ties with them. The site also calls for and facilitates employment repercussions, demanding that academic institutions should "take a far closer look at their roster of professors" and providing contact details for the departmental chairs of the professors concerned (Rate my racist professor 2020).

Such actions certainly do not apply due process in seeking to exact a punishment, and fail to show basic equal respect to that extent. But more fundamentally, they fail in equal respect because they purport to exact punishment according to judgements of a person's moral character. Where lists of offenders, open letters, and so on, call for admissions of guilt, not of an action but of a certain character, and demand humiliating apologies, these too have the effect of stigmatising the person's moral character as deviant, albeit implicitly. In the "Rebecca Tuvel controversy", an open letter called for the retraction of a peer-reviewed philosophy article, "In Defense of Transracialism", on the basis of alleged "failures of scholarship". Whilst putatively being about the credibility of the methodology, citation practices, and so on, the letter also accused Tuvel of causing "harms" to the vulnerable groups whose identities the article discussed, which it hinted were derived from discriminatory attitudes (Springer et al. 2017). Other academics and journalists responded by pointing out the many inaccuracies in the letter (Singal 2017). In cases such as the Tuvel letter, where the focus of the claims are on the actions of a person, it might not be OPS but still be morally faulty by seeking to exact personal costs, in the form of reputational costs, without due process, adequate opportunity for defence, or in fact veracity. They are a type of extra-judicial punishment, where the punishment includes reputational (and associated) social costs. Individuals valuing basic respect for others as a constraint on social interaction would refrain from performing, amplifying, or supporting such

actions. Attacking views as bigoted, providing reasoned argument as to why someone's perspective seems to be troubling, or criticising specific behaviour are not examples of OPS.

Individuals should, as a matter of ethics, avoid engaging in OPS for the reasons we have given. But should there be regulation of such acts? There are of course many forms of speech, such as some forms of lying, that are wrongful but are not legally prohibited, and for good reasons relating to people having rights to pursue their own lives and make their own moral mistakes. A certain threshold is needed for appropriate regulation. Clearly defamation, cyber-harassment, threats to personal safety and doxing should not be protected, and there are anyway existing remedies for these available in the civil and criminal law. The problem with OPS is that regulating it can clash significantly with free speech rights, and that drawing the line between comment and participating in shaming is difficult where no instrument of shaming (e.g. a collective letter or petition) is involved. However, where such overt instruments are employed, there seems to be no reason why individuals should not have remedies at their disposal.

Our primary proposal is that individuals who have been subject to a shaming should be given access to prominent "correction" or "right to reply" services, to balance out future judgments of their character. On Twitter, for instance, this might take the form of a "pinned" tweet from the victim on a shaming hashtag. Such a remedy restores to the victim the standing of a participant in public discourse, rather than an object being acted upon, restoring their self-respect and counteracting the Millian concern we have highlighted with the arbitrariness of public judgment. Social media platforms should also introduce provisions against OPS in their codes of conduct, which should be featured prominently on the platform, rather than being buried in lengthy terms and conditions, which few users end up reading. Furthermore, platforms should facilitate user-based feedback, trained

moderators, and algorithms, as Twitter, Facebook, and YouTube already do with other types of wrongful conduct.[24] These can act quickly where punitive shaming instruments are in play to stop the postings and block or remove the instruments.

Given that shaming can impact employment relations, regulation can be reformed to protect people from employers discriminating against them merely because they have been shamed online.[25] Using OPS-generated reputational effects as a ground for dismissal introduces moral character judgement as a criterion for employment, regardless of the person's ability to do her job well, adding to the injury of public ostracism. It disrespects the individual by reducing them to a public caricature. At the very least, employers could be legally required to be transparent regarding the information they gather on candidates and the criteria used in hiring and firing, and to make this explicit, up front, in the form of a social media exposure and reputation policy. This would allow for responsible policy-making on the appropriateness of such decision-making criteria.[26] There is also a strong case for responsible employers, universities, schools, academic journals, and membership organisations, adopting commitments and policies that they will not to bypass standard grievance and hiring procedures when they come under pressure from online mobbing. This would help ensure due process is followed and prevent knee-jerk firings, disciplinary measures and retractions. This would extend to university declarations against shamed individuals ("distancing themselves from them") where the person has not breached any code or law.

Conclusion

[24] See, e.g. Facebook 2020.

[25] According to one estimate, 90% of employers search their candidates online record with some using sophisticated tools to search for any traces of controversy, no matter how well hidden (Citron 2014: 8).

[26] In the case of Professor Steven Salaita, his appointment to the University of Illinois was rescinded following tweets critical of Israeli military action (Associated Press 2015).

The above principles can act as a reference point for identifying, naming, and rejecting OPS where one sees it. They can give victims an ethical vocabulary in which to articulate the wrong of what is being done to them – the informal punishment and the consequences of it being exacted. Individuals can also use these standards to critique the actions of those using such activities as a *modus operandi*. Digital technology and social media have given rise to new power relations between individuals and informal collectives which pose new questions of accountability. Online public shaming is one of the most ethically challenging products of those transformations. We have argued it is ethically wrong, on recognisable moral grounds, wronging its targets in two key respects: breaching basic respect and imposing informal punishments that are inherently not amenable to due process. Its consequences can be severe and unconstrained. To act ethically, on our analysis, users of social media should forego such activities, while social media platforms, employers, universities and other organisations, should take responsible action to address the relevant reputational effects. Social ostracism in the form of public shaming is an illiberal form of social regulation and so is its digital analogue.

Dr. Guy Aitchison (Loughborough University) g.aitchison@lboro.ac.uk

Dr. Saladin Meckled-Garcia (University College London) s.meckled-garcia@ucl.ac.uk

References

Acton, H. B. ed. 1969. *The Philosophy of Punishment*, London: MacMillan.

Associated Press. 2015. "University of Illinois censured for pulling Steven Salaita job over anti-Israel tweets", *The Guardian*, 14 June, Available at: https://www.theguardian.com/us-news/2015/jun/14/university-of-illinois-censured-for-pulling-steven-salaita-job-over-anti-israel-tweets. Accessed: May 20, 2017.

Archard, David. 2014. "Insults, free speech and offensiveness." *Journal of Applied Philosophy, 31*(2), pp.127-141.

Austin, John Langshaw. 1975, *How to do things with words*. Vol. 88. Oxford: Oxford University Press.

Badash, Nadeem. 2018. "Facebook to contact 87 million users affected by data breach", *The Guardian*, April 8 , Available at: https://www.theguardian.com/technology/2018/apr/08/facebook-to-contact-the-87-million-users-affected-by-data-breach Accessed: January 31.

Bicchieri, Cristina. 2016. *Norms in the Wild: How to Diagnose, Measure and Change Social Norms*, Oxford: Oxford University Press.

Billingham, Paul and Tom Parr. 2019. "Online Public Shaming: Virtues and Vices." *Journal of Social Philosophy* December 1. https://doi.org/10.1111/josp.12308

Brennan, Geoffrey, Lina Eriksson, Robert E. Goodin and Nicholas Southwood. 2013. *Explaining Norms*, Oxford: Oxford University Press.

Citron, Danielle Keats. 2014. *Hate crimes in cyberspace*. Harvard University Press.

Citron, Danielle Keats and Mary Anne Franks. 2014. "Criminalizing revenge porn." *Wake Forest Law Review*, 49: 345.

Council of Europe. 2013. *Guide on Article 6 of the European Convention on Human Rights*, Strasbourg: Council of Europe.

Dagger, Richard. 2011. "Social Contracts, Fair Play, and the Justification of Punishment", *Ohio State Journal of Criminal Law*, 8.

Darwall, Stephen L. 1977. "Two kinds of respect." *Ethics* 88 (1), 36-49.

Deonna, Julien A., Deonna, J. A., Raffaelle Rodogno, & Fabrice Teroni. 2011. *In Defense of Shame: The Faces of an Emotion*. New York: Oxford University Press.

Duff, Richard. A. 2001. *Punishment, Communication, and Community*, New York: Oxford University Press, USA.

Duff, Richard. A. 1991. *Trials and Punishments*. Cambridge University Press.

Duchen, Claire. 2000. "Crime and punishment in liberated France, the case of the *femmes tondues*", in Claire Duchen, and Irene Bandhauer-Schöffman, *When the War was Over: Women, War and Peace in Europe, 1940-1956*, Leicester University Press.

Flew, Antony. 1969. "The justification of punishment" in Acton, H. B. ed., *The Philosophy of Punishment*, Macmillan: London.

Facebook. 2019. Facebook's "Community Standards", Available at https://www.facebook.com/communitystandards Accessed April 10, 2019.

Feinberg, Joel. 1965. "The expressive function of punishment." *The Monist*, 397-423.

Fiegerman, Seth. 2018. "Twitter is profitable again and adding users", *CNN.com*, April 25, Available at: https://money.cnn.com/2018/04/25/technology/twitter-earnings/index.html Accessed: January 31, 2019.

Foreman, Jonathan. 2015. "The Timothy Hunt Witch Hunt", *Commentary Magazine*, September 2015, Available at:https://www.commentarymagazine.com/articles/the-timothy-hunt-witch-hunt/ Accessed: April 12, 2017.

Forsdyke, Sara L. 2005. *Exile, Ostracism, and Democracy: The Politics of Expulsion in Ancient Greece*, Princeton: Princeton University Press.

Frost, Chris. 2015. *Journalism Ethics and Regulation*, Abingdon: Routledge.

Goldman, Lauren M. 2015. "Trending now: the use of social media websites in public shaming punishments." *American Criminal Law Review* 52, 415.

Goodin, Robert E. 2019. "Rough justice" *Jus Cogens*, 1 (1) pp.77-96.

Hovellin" Hermit. 2018. @HovellingHermit, Twitter, 10:55am, October 23, At: https://twitter.com/HovellingHermit/status/1054687676354519040

Jacquet, Jennifer. 2016. *Is Shame Necessary?: New Uses for an Old Tool*. New York: Vintage.

Khomami, Nadia. 2017. "#MeToo: how a hashtag became a rallying cry against sexual harassment", *The Guardian*, 20 October, At:https://www.theguardian.com/world/2017/oct/20/women-worldwide-use-hashtag-metoo-against-sexual-harassment Accessed at: May 16, 2020.

Klonick, Kate. 2015. "Re-shaming the debate: social norms, shame, and regulation in an internet age." *Modern Law Review*. 75: 1029.

Kutz, Christopher. 2000. *Complicity*, Cambridge: Cambridge University Press.

Laidlaw, Emily B. 2017. "Online shaming and the right to privacy." *Laws* 6.1: 3.

Larmore, Charles. 1989. "Liberal neutrality", *Political Theory*, *17*(4), pp.580-581.

LBC. 2016. "Sexism shame scientist considered suicide", 28 January, LBC.co.uk, Available at: https://www.lbc.co.uk/hot-topics/everyday-sexism/sexism-shame-scientist-considered-suicide-124010/, Accessed: April 12, 2019.

Levmore, Saul, and Martha C. Nussbaum. 2010. eds. *The Offensive Internet.* Harvard University Press.

Massaro, Toni M. 1997. "The meanings of shame: Implications for legal reform." *Psychology, Public Policy, and Law, 3*(4), p.645.

McDonald, Soraya Nadia. 2014."Racists Getting Fired" exposes weaknesses of Internet vigilantism, no matter how well-intentioned", *Washington Post*, 2 December, Available at: https://www.washingtonpost.com/news/morning-mix/wp/2014/12/02/racists-getting-fired-exposes-weaknesses-of-internet-vigilantism-no-matter-how-well-intentioned/?utm_term=.2c1e01bfedb6 Accessed: January 31, 2019.

Meckled-Garcia, Saladin. 2017. "On the object and scope of neutrality", in Laborde, C. and Bardon, A., eds., *Religion in Liberal Political Philosophy*, Oxford: Oxford University Press.

Norlock, Kathryn J. 2017. "Online Shaming." *Social Philosophy Today* 33: 187-197.

Mill, John Stuart. 2003. *On Liberty. Rethinking the Western Tradition*, eds. David Bromwich, and George Kateb, Yale University Press.

Nussbaum, Martha C. 2004. *Hiding from Humanity: Disgust, Shame, and the Law*, Princeton University Press.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford Oxford University Press.

Rate my Racist Professor, Ratemyracistprofessor.com Available at: https://ratemyracistprofessor.com/ Accessed: March 8, 2020.

Rolph, David. 2013. "Defamation by Social Media", *Precedent*, Vol. 117, pp. 16-21.

Ronson, Jon. 2015. *So you've been Publicly Shamed*. Riverhead Books Hardcover.

Schauer, Frederick. 2015."Free Speech on Tuesdays", *Law & Philosophy, 34*(2), pp.119-140.

Simmons, A. John. 1995. "Locke and the right to punish" in John A. Simmons, A. Marshall Cohen, Joshua Cohen, Charles R. Beitz eds., *Punishment*, Princeton: Princeton UP.

Singal, Jesse. 2015. "Monica Foy, the Victim of a Terrifying Right-Wing Internet-Shaming, Speaks Out", *NY Mag*, September 4, Available at: http://nymag.com/intelligencer/2015/09/victim-of-a-scary-web-shaming-speaks-out.html?gtm=bottom&gtm=top Accessed: January 31, 2019.

Singal, Jesse. 2017. "This Is What a Modern-Day Witch Hunt Looks Like", *NY Mag*, May 2, Available at: http://nymag.com/intelligencer/2017/05/transracialism-article-controversy.html Accessed: April 12, 2019.

Salaita, Steven. 2019. "An honest living", *Steven Salaita blog*, February 17 2019, Available at: https://stevesalaita.com/an-honest-living/ Accessed: April 12, 2020.

Smith, Patrick. 2015. ”This Is What It"s Like To Go To Prison For Trolling", *BuzzFeed*, March 2, Available at: https://www.buzzfeed.com/patricksmith/isabella-sorley-john-nimmo-interview?utm_term=.kblmgJLaz#.eh2ZxgNpB Accessed: May 16, 2020.

Solove, Daniel J. 2007. *The Future of Reputation: Gossip, Rumor, and Privacy on the Internet* New Haven: Yale University Press.

Solove, Daniel J. 2010. "Speech, privacy, and reputation on the Internet", in Saul Levmore and Martha Nussbaum, eds. *The Offensive Internet.* Harvard University Press, pp.15-30.

Springer, et al. 2017. “Open letter to Hypatia”, *Archive.today,* Available at: https://archive.is/lUeR4 Accessed: March 9, 2020.

Suler, John. 2004. "The online disinhibition effect." *Cyberpsychology & Behavior* 7.3: 321-326.

Sunstein, Cass R. 2010. "Believing False Rumors," in Saul Levmore and Martha C. Nussbaum. *eds. The Offensive Internet.* Harvard University Press, pp. 91–106.

Taylor, Gabriele. 1985. *Pride, Shame, and Guilt: Emotions of Self-Assessment*, Oxford, Clarendon Press.

Thomason, Krista K. 2018. *Naked: The Dark Side of Shame and Moral life*. New York: Oxford University Press.

Twitter, "The Twitter Rules", Available at: https://help.twitter.com/en/rules-and-policies/twitter-rules Accessed: April 10, 2010.

Uhl, Carolyn A. 2018 et al. "An examination of nonconsensual pornography websites." *Feminism & Psychology*, 28.1: 50-68.

Velleman, J. David. 2001. ”The genesis of shame." *Philosophy & Public Affairs* 30.1: 27-52.

Waldron, Jeremy. 1981. "A right to do wrong." *Ethics*, 92.1: 21-39.

Whitman, James Q. 1997. "What is wrong with inflicting shame sanctions." *Yale Law Journal* 107: 1055.

Wood, Graeme. 2013. "Scrubbed", *NY Mag*, June 14, Available at: http://nymag.com/nymag/features/online-reputation-management-2013-6, Accessed May 28, 2017.