The many shades of digital vigilantism. A typology of online self-justice

Benjamin Loveluck

Citation: Loveluck, B. (2020), "The many shades of digital vigilantism.
A typology of online self-justice", *Global Crime* vol. 21 n° 3-4, pp. 213-241

The many shades of digital vigilantism.
A typology of online self-justice

Benjamin Loveluck, i3-SES, Telecom Paris

Abstract

Digital vigilantism involves direct online actions of targeted surveillance, dissuasion or punishment which tend to rely on public denunciation or on an excess of unsolicited attention, and are carried out in the name of justice, order or safety. Drawing on a diversity of case studies, this article seeks to provide a comprehensive picture of its manifestations, addressing both the social practices and digital media dynamics involved. It presents a typology which distinguishes between four ideal types of digital vigilantism: flagging, investigating, hounding, and organised leaking.

Keywords:

Digital vigilantism; Surveillance; Online denunciation; Typology

Introduction

The internet has long been presented as encouraging participation, a more active and decentralized public sphere, and an autonomous space of citizen power, premised on a fairly civil understanding of democracy[1]. It has also been hailed as the triumph of 'mass self-communication' in the service of political progress[2]. However, as recent developments have made clear, it is also the place of messier and more problematic forms of engagement. This is increasingly obvious since hate speech, discrimination and harassment have been shown

to figure prominently in online interactions[3]. A category of practices, however, straddles the line between these two extremes, since they conform neither to participative ideals nor to due process of policing and law enforcement, but are nonetheless undertaken in the name of justice, order or safety, and may to some extent be considered as a mode of political participation as well as a form of moral regulation and a response to criminal activities.

Circulating pictures of men spreading their legs wide apart when sitting in a public transport, in order to denounce this type of behaviour, and leading to regular contributions under the #manspreading hashtag. Attempting to solve a mystery (such as a missing person) or identify the perpetrator of a crime (such as a thief or burglar caught on camera) by sharing evidence via social media such as Facebook or Twitter or even on dedicated websites, thus initiating crowdsourced efforts at investigating the matter. 'Naming and shaming' on social media someone accused of wrongdoing, or whose legal punishment is considered insufficient. Deploying efforts at uncovering the identity, collecting visual evidence and publishing personal information of individuals accused of animal cruelty, and encouraging campaigns to damage their reputation or to harass them via email, text messages or social media. Setting up an application enabling researchers to anonymously denounce forged results or biases in scientific publications. All these situations involve direct forms of intervention online, targeting individuals, their behaviour or organisations in order to deter or punish them outside of institutional frameworks and accepted norms of 'civility'. They can therefore be referred to as instances of *digital vigilantism*, whereby individuals seem to be 'taking the law into their own hands' online.

The aim of this article is to provide a more nuanced understanding of such practices, and thus contribute to clarifying the conceptual contours of digital vigilantism. In order to do so, a range of case studies have been collected and characterized, drawing on analytical concepts derived from both the sociology of disputes and contentious politics studies. Key dimensions were identified in order to distinguish between different forms of digital vigilantism, and the resulting typology includes four ideal types: flagging, investigating, hounding and organised leaking[4].

Approaching digital vigilantism

Despite many attempts at providing conceptual clarification, vigilantism in general is notoriously difficult to define[5]. The establishment of formal criteria – regarding the degree of violence and illegality involved, the relationship with official law enforcement (collaboration or defiance), its collective or individual dimensions, its planned or

spontaneous character, and even its conservative or progressive nature – has proved challenging.

Political scientists and theorists have considered vigilantism primarily as the (usually violent) expression of a collective power seeking to assert or to restore order through direct punishment, in defiance of existing legal and institutional frameworks. It has therefore been classically referred to as a form of 'establishment violence'[6], and more recently as an expression of popular sovereignty as 'uncivil disobedience'[7] for whom the 'rule of the people' trumps the 'rule of the law'. Criminologists on the other hand, following the work of Les Johnston[8], have pointed out that some vigilante actions may remain within the boundaries of the law and may not involve a punishment. They have therefore avoided taking the legal or the punitive dimensions as decisive criteria, and have tended instead to address vigilantism as an organised and 'forceful' reaction to either criminal or social/moral transgressions on the part of volunteer citizens, whose objective is to ensure the security of an established order. From both a political science and a criminology perspective however, vigilantism questions the relation to the state and its institutions – its legal framework, its judicial system and its police forces. It can be understood as a weakened capacity or resolution on the part of the state to exert its sovereignty in certain instances, as a challenge to its prerogatives by 'concerned citizens' willing to take action, or even as a form of delegation or outsourcing of these prerogatives to private parties in line with a neoliberal reorganisation of institutions.

Digital vigilantism involves online direct actions in response to perceived civil or moral transgressions, crimes or injustices. It complicates matters even further for at least two main reasons. The first one is that it is less clearly associated with violence understood as physical coercion, insofar as the bulk of these online actions take the form of public denunciations or targeted harassment – sometimes to inflict direct damage but more often to identify, humiliate or disgrace the perpetrators, or to trigger a response from the authorities and the judicial system. It thus relies primarily on 'unwanted', 'intense' and 'enduring' forms of visibility[9]. Although this can be considered a form of symbolic violence, which can entail very real psychological, social, and sometimes material and even physical consequences, its punitive dimensions are often less straightforward. The second reason is that the internet has made such initiatives more readily available and even commonplace, by lowering the bar for a wide range of direct actions and by commonly relying on self-regulation in order to police online interactions.

These specificities have led an author such as Finn Brunton to discard the notion of vigilantism altogether in a digital context, and to draw parallels instead with intense forms of

public humiliation and harassment such as the age-old tradition of *charivari* and 'rough music' which used to be directed against transgressions of social mores (e.g. illegitimate couples):

> What we are discussing here is a complex political performance that is built out of mocking laughter, insults, masking and anonymity, and the mingling between active crowds and passive audiences. (...) The charivari, both on- and offline, from the July Monarchy to antispam vitriol and 4chan's lulz-driven crusades in the present day, draws much of its efficacy from renegotiating the boundaries between public and private life.[10]

This perspective rightfully draws attention to the issue of shame and humiliation as a collapse of the distinction between private and public matters, and as a key resource for collective regulation of behaviours. However, it does not address the whole range of public denunciations online and their punitive implications. It also tends to downplay the efforts which can be channelled in investigating a given situation online and disciplining perceived offenders, in many cases the sustained nature of such actions, and the very serious consequences they can entail in terms of damage to an individual's reputation or dealing with sometimes vicious forms of harassment.

More generally, digital vigilantism hinges on two important characteristics of digital media. The first one is the *regime of visibility* involved. In a trivial sense, this is directly related to the ease with which information, data and especially photos and videos can be recorded, copied, shared and widely published online. Moreover, as is now well known, online social practices commonly involve the logging and archiving of one's activities (on forums, websites, social media platforms etc.) as well as a marked tendency towards selective exposure of the self on the Web. The digital era has thus seen the rise of an environment of mutual or 'lateral' surveillance between individuals monitoring each other[11], which both deepens and extends previous forms of collective observation and social control of individual behaviour. Social validation now rests on a subtle balance between self-exposure and concealment of privacy, which can topple suddenly and unexpectedly when private information is given unwanted publicity or is repurposed and shared with people it wasn't initially intended for (a phenomenon known as 'context collapse'[12]).

The second characteristic is the *culture of self-regulation* which dominates online practices, and the governance mechanisms of online communities based on voluntary participation[13]. This includes more or less formalised norms of expected behaviour initially found on newsgroups, mailing lists or forums, enforced by moderators and admins[14], as well as more sophisticated forms of conflict resolution within Wikipedia[15], all the way to reporting tools and flagging systems deployed on social media platforms to ensure users abide by the terms of use[16]. These are driven by an ideal of immanent organisation of interactions and

exchanges as well as a defiance towards state institutions[17], which are conducive to privileging more horizontal and person-to-person forms of conflict resolution – including direct retaliation.

This research thus relies on a broad definition of digital vigilantism as *direct online actions of targeted surveillance, dissuasion or punishment which tend to rely on public denunciation or an excess of unsolicited attention, and are carried out in the name of justice, order or safety*. I argue that digital vigilantism so understood is a valid notion, but that the many nuances in its effective manifestations must be precisely accounted for – including borderline forms at both ends of the spectrum: those which may appear 'milder' on the one hand, as well as those which can be more consequential on the other.

A typology of online self-justice

Methodology

The research adopts a comparative perspective, based on a total of about fifty different cases which are within the scope of digital vigilantism as defined above[18], to ensure a sufficient diversity of situations. Most of the cases were collected first hand over a period of several years (2015-2018), mainly from France and the UK while additional, already contextualised cases from the US, China or Russia were drawn from secondary sources and relevant academic literature (also providing conceptual contributions which are included in the discussion). Collection of empirical material involved mixed methods of digital ethnography[19] based on participant observations, regular archiving of data from digital platforms or websites (using both screenshots and Web or social media scraping when possible)[20], and content analysis of Web pages, forums, IRCs, chat apps and social media.

The cases were documented with varying degrees of detail, but sufficiently to establish a set of differentiating criteria for comparison. No attempt was made to quantify these differences and measure any causal properties however. Rather, the aim was to provide a heuristic framework by identifying contrasting courses of action, which can be understood as 'ideal types' or 'pure types' of digital vigilantism[21]. In order to determine meaningful criteria for interpreting the data and building these ideal types, two theoretical approaches were combined which allowed to account for both the *moral* and the *political* aspects involved in digital vigilantism.

The first set of criteria sought to characterize how situations of conflict are assessed by actors as well as the types of reaction they generate, and is derived from the *sociology of disputes*[22] and the 'sociology of scandals'[23]. As a pragmatic sociology which focuses on

conflicts and controversies as performative processes, the sociology of disputes is particularly adapted to the object at hand. Indeed, it takes seriously the feeling of outrage which triggers such actions and the ordinary '*sense of justice*' of the actors involved. Moreover, it focuses on *public denunciation* and its consequences, as an expression of this outrage and as an empirical object of study. It involves tracing the dynamics of a controversy along with the situated judgments it generates, by relying mainly on the justifications provided by actors in order to explain why they did what they did, or why they criticize a particular behaviour. It is therefore particularly relevant for studying digital vigilantism, which often involves forms of public denunciation and shaming. Drawing on such a 'grammar' of moral action, three key dimensions were singled out: the *triggers* of a given action (the cause of the outrage, from uncivil behaviour and offences all the way to crimes and systemic corruption), the nature of its *targets* (whether directed towards a behaviour, a person, or an institution), and the *motives* which sustain it (the effects which are sought – deterrence, identification, punishment or systemic change).

The second set of criteria sought to characterize the types of action taken to resolve the conflict, and relies on methods and insights derived from *contentious politics studies*[24]. The latter normally focuses on political conflicts, violence and social movements. One of its key analytical tool consists in identifying action repertoires – a set of resources and routinized courses of action which actors can resort to in order to voice a claim (such as petitions, demonstrations, strikes, civil disobedience or sabotage), and which vary in time and space according to the context of a given situation, the relation to the authorities and the intentions of the actors[25]. In the case of digital vigilantism, the opportunities and constraints of the social and political environment are supplemented by the notion of digital affordances[26] and internet-related action repertoires[27]. Two main components of action repertoires were mobilised to characterize the different types of digital vigilantism: the *tactics* involved on one hand (a chosen course of action such as naming, shaming, raiding or incentivizing information disclosure) and the *organizational forms* adopted on the other (modes of coordination in the event of a collective action, which range from ad-hoc and loosely coordinated activities to setting up dedicated resources, leveraging pre-existing networks and engaging in sustained and rehearsed collective efforts).

These five key dimensions – triggers, targets, motives, tactics and organizational forms – were assessed qualitatively for each case, bearing in mind that a certain degree of interrelation exists between them. Taken together, they provided lenses through which to interpret the diversity of cases at hand, and enabled the development of a rich comparative

and relational perspective, which accounts for both social practices and digital media dynamics and which qualifies the form and degree of denunciation, shaming or harassment taking place (see Table 1 for a summary view). Finally, four ideal types were identified, which are further described and analysed in the following sections: *flagging*, *investigating*, *hounding* and *organised leaking*.

Table 1. A typology of digital vigilantism

		Flagging	**Investigating**	**Hounding**	**Organised leaking**
Trigger		Breach of social norms	From minor offences to crimes & terrorist actions		From uncivil behaviour to fraud & systemic corruption
Target		Behaviour or group	Person or category of person		Institution or organisation
Motive		Public objection to a type of behaviour	Identification of suspect	Direct punishment of accused & intimidation	Systemic change or improvement
Main action repertoires	*Tactics*	Shaming	Naming	From naming & shaming to doxing & raiding/swarming	Incentivized & systematized disclosure of information
	Organizational form	From ad-hoc & loose coordination (e.g. via hashtags) to specific websites or social media pages & leveraging of pre-existing networks		From ad-hoc & loose coordination to sustained & rehearsed actions via specific communication platforms	Long-term collective effort federated via specially tailored procedures & tools
Examples		Circulating (anonymized) pictures of bad drivers or ill-mannered passengers in public transport	Sharing pictures of suspected thieves or vandals as evidence; websleuthing	From sharing pictures of individuals accused of wrongdoing to hunting down animal abusers or paedophiles; Anonymous targeted 'ops'	WikiLeaks, PubPeer

Flagging

Flagging covers a great variety of cases which are generally of low intensity, and which involve *shaming a behaviour*. It leverages the affordances of social media to alert users about different

categories of actions or conducts which are considered uncivil and a breach of social norms. This usually takes the form of public shaming, often by circulating pictures of a behaviour deemed to be objectionable – but avoiding to target the specific persons involved. The actions are generally loosely coordinated, although they can be aggregated and coordinated via specific hashtags, and sometimes webpages or social media pages.

One typical focus of flagging involves traffic incidents, such as 'bad drivers' or 'badly parked cars'. For instance in France, pictures of cars parked in a dangerous way – particularly when they are blocking a bike lane – may be circulated via Twitter or Instagram with a dedicated hashtag such as #GCUM (for '*garé comme une merde*' i.e. 'parked like shit'). A website has also been set up (http://www.garecommeunemerde.fr) where stickers can be bought, to be placed on the offending vehicle. Both forms of shaming – singling out the behaviour to passers-by as well as to a wider audience on social media – are presented as constructive ways of venting one's anger. Generally, the drivers cannot be recognized and the licence plates are concealed, though in some cases they are left on the picture, and furthermore the authorities (the local police or the mayor) can be added to the loop, for instance by mentioning their Twitter account in a tweet. In a number of cases, cyclists have even decided to install cameras in order to film their experience on the road, not unlike dashcams set up by car drivers in certain countries in case of accidents where visual evidence could be needed[28]. These videos can then find their way to Facebook or YouTube where they are used to attract attention on dangerous situations[29], and also in some cases as a way to protect oneself from hostility and sometimes physical violence on the part of drivers whose behaviour has been pointed out.

Other forms of behaviour deemed uncivil or shameful can also be flagged. *Passenger Shaming* is a highly successful Facebook page and associated Instagram account[30] dedicated to shaming plane passengers who take their socks off, leave their garbage behind, or take up too much space, by circulating crowdsourced pictures and videos. Set up by a former flight attendant who initially held a blog ('Rants of a sassy stew'), it has even become a brand which can monetize the traffic it generates and also sells a range of associated merchandising. Many initiatives also focus on gender issues in the public sphere – for instance 'manspreading' in public transports mentioned in the introduction (figure 1), or 'catcalling' in the street – and have sometimes attracted considerable media attention.

Figure 1. A picture taken on a UK bus and shared on Twitter, captured 16 November 2018.

In all these cases, the individuals who are shamed in such situations are generally not recognizable: their face is outside the frame or cropped out, very little elements of context are provided, they are normally unaware they have been caught on camera, and the circulation of the images is limited. The aim therefore is not to identify them directly, or to confront them with the recorded behaviour. However, from the comments that are generated by the shared images, it appears that some degree of satisfaction is derived from exposing behaviours which may be considered either 'unacceptable' or 'deviant', even though their authors are oblivious to what is happening. The (anonymized) humiliation still constitutes a form of mild punishment in the eye of the beholder, as well as sometimes an assertion of moral superiority by making fun of these attitudes. Due to this anonymity, the consequences for the persons involved are minimal, and the flagging only generates a small amount of controversy if at all. Rather than a targeted assault, it represents an impersonal call to change an attitude or to fall back in line, a mild threat which can also often be associated with a form of humour or sarcasm as the blameworthy behaviour is derided.

There are cases however, where flagging can take greater proportions – particularly when at least one of two conditions are met: 1) the circulation of shamed behaviour is sustained and within a local community, and/or 2) mainstream media pick up on the event and amplify its effects. For instance, a public Facebook page was set up anonymously in March 2016 by citizens in the small town of Étaples in Northern France, called '*Incivilités étaploises*' ('Incivilities in Étaples'). Its stated purpose is to 'document all the environmental,

hygienic and regulatory incivilities committed in our town'[31]. Only about 1,500 people follow the page (number of 'likes' on October 28, 2018), but the overall population of Étaples is about 11,200.

Analysis of the page over a two-year span shows that, despite periods of lesser activity, the posts shared on the page (predominantly by the page owners) often trigger a fair number of reactions and comments – with peaks between 20 and 70 comments. The data allows us to track the publications which generate the most reactions and assess the nature of the interactions. It appears that contributors generally vent their anger against alleged examples of incivility or insecurity, sometimes in quite crude terms, sometimes criticizing the local authorities for their lack of reaction. On the other hand, users who either follow this page or have just discovered it can be quite critical of the initiative, or deride it and make fun of the page.

Figure 2. 'Incivilités étaploises' Facebook page, captured 11 September 2017. Status translation: 'Unacceptable and intolerable behaviour!!'.

A considerable spike can be identified on September 8, 2017, when a burst of vandalism led to many damages in the streets and on vehicles (figure 2) for which three young men were later arrested by the police). Most of the publications however deal with litter in the streets, fouling the pavement, vandalised public or private property, noise nuisances, and badly parked cars. In some cases however, and due to the local nature of the page, users may

recognize their own or their neighbours' property (such as a car or a vandalized space) and attempt to justify their actions, thus leading to heated discussions in the comments (figure 3). Some users admit to following the page only in order to check if anything is being said about their vicinity or themselves.

Figure 3. 'Incivilités étaploises' Facebook page, captured 10 July 2017. Status translation: 'A woman from Étaples has sent us this photo of an unauthorised vehicle parked on the disabled space. Indeed, nothing justifies nor authorises such behaviour, especially when there are free parking spaces just a few meters away!! Showing a good example should be part of certain functions!! [the vehicle belongs to town hall employees who are doing repair work]'.

It must also be mentioned that the regional daily newspaper *La Voix du nord* ('The Voice of the North') attracted attention to the page not long after it was created, by publishing an article about it[32]. The anonymity of its authors as well as the method of public denunciation were considered questionable – quoting the local authorities for whom the move amounted to '*délation*' (a form of reporting and informing on others which, in France,

is usually associated with collaborators of the Nazi occupation during World War II) but also stressing, however, that license plates are blurred and that no names are published. As is therefore clear from this example, the effects of flagging may be amplified if they are localised, and/or if they are picked up by mainstream media. It illustrates how drawing attention to a behaviour (or its consequences) can slip towards becoming suspicious of specific individuals or groups of people, and sometimes directly identifying them.

Flagging can thus involve a form of community policing when targeting, on a given geographic area, specific behaviours which are considered offensive or dangerous. A Bordeaux neighbourhood set up a Facebook page in 2013 where users were encouraged to take part in a 'deal safari' by taking pictures of drug dealers in their street and posting them on the page. This was intended to intimidate the drug dealers directly as well as to publicize the issue and thus attract police – as well as media – attention. Here too the people involved had to deny being *délateurs* in interviews given to the press, a criticism which was anticipated on their Facebook page and on the posters displayed in the streets which stated that 'the neighbourhood is upset not fascist' ('*les habitants du quartier sont fâchés pas fachos*'). Indeed, this shows that such actions are inherently controversial and that their authors may be aware they are themselves engaging in a form of transgression of established collective norms. These are heavily dependent on the situation however, with public scrutiny taking different forms according to the role ascribed to institutions – levels of legitimacy and trust but also political and cultural context (civilian policing and neighbourhood watches, for instance, being much more readily accepted throughout the US or in Britain since the 1980s). In the Bordeaux case although the page was quickly taken down, it enabled the identification of a suspect and eventually led to a conviction[33], thus edging towards the next category in this typology.

Investigating

Investigating aims at *naming a person*. Here, a collective effort is made to identify individuals suspected of wrongdoings, which can range from minor offences such as theft, all the way to crimes and terrorist activities. This usually involves a call to the public after the event has occurred and sometimes as it unfolds, with personal digital traces and records (such as pictures of a suspect or a crime scene, screenshots from social media accounts, or contributions on forums and chats) being collected as evidence. The material is then shared directly via dedicated platforms or on social media in order to leverage 'collective intelligence'. The aim is to solve a puzzle, but also to identify the persons involved. The initiative can come from 'concerned individuals', who in some cases may share their findings

with the authorities – but it can also be prompted by the authorities themselves, for instance when launching a 'call to witnesses', which is in line with more common forms of cooperation between police and public. Beyond any questions raised about the legitimate ways for citizens to collaborate with the police, online investigations are therefore inherently ambiguous because public exposure *in itself* already constitutes a form of punishment – through potential shaming, which can also sometimes lead to retaliatory actions of harassment.

Investigating is illustrated by a long-time phenomenon called 'websleuthing', which involves 'varying levels of amateur detective work including but not limited to searching for information, uploading documents, images and videos, commenting, debating, theorising, analysing, identifying suspects and attempting to engage with law enforcement and other organisations and individuals connected to the cases'[34]. It is particularly common in the US, where case materials can more easily be accessed through public information laws in certain states. Beyond the fact that it is led by amateurs outside formal investigations procedures, one key aspect of websleuthing is that it is held publicly: it usually hinges on the affordances of online forums such as dedicated websites (e.g. Websleuths.com which was launched in 1999 and claims nearly 138,000 members, over 300,000 threads and 13,5 million messages as of July 2018) or subsections of popular platforms such as Reddit (e.g. 'Unresolved Mysteries' and the 'Reddit Bureau of Investigations' – RBI)[35]. Users first create a post – often asking for help to identify something or someone, for advice in finding information, or just to share a 'mystery' they have come across. Other users then contribute to the different threads by providing suggestions, recommendations and sometimes technical expertise.

Yardley et al. stress that websleuthing has wider cultural roots. These include the longstanding role played by the media in spreading representations of crime, feeding curiosity for solving crimes as 'infotainment' and encouraging discussion and speculation on the cases at hand, as well as the more recent 'participatory culture' trope associated with the affordances of networked media. Examination of motives as reported in the press showed a predominance of wanting to achieve justice or closure, although the challenge of solving a puzzle and various forms of fascination were also mentioned, with some contributors reporting a sense of 'duty' to carry out such investigations; however despite some high-profile success stories, effectiveness of websleuthing appears limited, and in some cases interferes with the police or judicial process, sometimes even impacting suspects or victims[36].

A prominent example of the limits of such crowdsourced investigation occurred after the Boston marathon bombings of 2013, mainly on the Reddit social platform[37], which led

to a completely misguided (and racially-tinged) identification of several innocent persons[38]. It was a turning point for Reddit, triggering official apologies from the owners of the website:

> A few years ago, reddit enacted a policy to not allow personal information on the site. This was because 'let's find out who this is' events frequently result in witch hunts, often incorrectly identifying innocent suspects and disrupting or ruining their lives. We hoped that the crowdsourced search for new information would not spark exactly this type of witch hunt. We were wrong. [39]

This contributed to a gradual reassessment of both its radical free speech and 'participatory' culture: as Adrienne Massanari argues, both the affordances of the platform and the shared culture of its members are conducive to such behaviours[40]. It should be noted however that the redditors' initiative was spurred by the FBI, which initially asked the public to submit any photographic evidence they might have. A lot of this material also found its way to Reddit or other social media and from there (along with speculations and rumours) to the mainstream media – such as the front page of the *New York Post*, falsely singling out two men of colour referred to as 'bag men' in capital letters. However, when the authorities did manage to identify the two suspects several days later, they had to release a picture much earlier than intended: although this was presented as a call to witnesses, it was done partly to prevent further speculation and misdirected targeting of suspects on social media and in the press.

Another illustration of the complex relationship between the authorities, the media and websleuths concerns the riots which followed a hockey match in Vancouver in 2011[41]. During and after the riots, many pictures of vandalism and violence were shared and commented on Facebook, sometimes by the offenders themselves. The police set up a dedicated website and encouraged the public not only to send them incriminating pictures recorded through their mobile phones or found on social media, but also to tag and identify any suspects where possible. They presented this as a form of 'civic action', while warning against temptations of 'vigilante justice'. Thousands of images and hours of video (totalling more than 30 terabytes of data) were sent, leading to dozens of arrests. The virulence of the reactions however, showed that the public's involvement went beyond mere cooperation with law enforcement:

> This is a prime example of crowd-sourced policing – the organization and use of everyday technology by citizens not affiliated with law enforcement – to scrutinize and persecute fellow citizens suspected of criminal behaviour. These efforts generated public criticism, because of the prejudicial fervour with which users identified and criminalized suspected rioters.[42]

The aim here was not merely to solve criminal cases, but also to punish the individuals involved and restore social order by tracking down 'deviants'. Once identified on line, some suspects (including minors) were harassed and threatened, others were sacked from their jobs and in one case a family had to move away.

Both previous examples involve a security crisis, however websleuthing more commonly takes place when prominent criminal cases remain unsolved. It can garner a fairly important number of participants via ad hoc means such as Facebook groups, forums, or dedicated websites. The digital environment becomes both a shared space to discuss the matter and to organise, and a resource where information can be gleaned. Other more mundane forms of investigation may also take place over social media. These involve the simple sharing of a message, for instance on Twitter or on Facebook, calling for help to identify the authors of petty crimes such as vandalism or theft. The message can include a picture or video of the suspects if caught on camera, or of the stolen object or vandalised place, and often presents a local character. In one example (figure 4), CCTV footage of two young men trying to break into a shop has been shared nearly 2,000 times over Facebook, mainly among people belonging to the same local borough in the North of England.

Figure 4. Facebook publication, 8 July 2018 (faces have been anonymised but are visible on the original source).

Such information serves as clues or as evidence, which are shared with the public rather than exclusively with the police, often in the belief that the crime isn't serious enough and will not warrant the attention of the authorities, but also in the hope that such a direct call to witnesses will be more effective. The comments show a good deal of empathising and encouragements for the victims, as well as copious amounts of swearing against the suspects ('scum' and 'scumbags' being the most common). They also show that in this case, the aim is primarily to collect information and perhaps find someone who can identify the culprits, since the message is mainly shared within a small community ('Saw these guys at the traffic lights near asda [supermarket]' says one comment; 'I woke to my father shouting at them' says another). Advice is also provided ('If u still have the stone used can they not fingerprint it?'). Finally, the victims are urged to contact the police, while some voices question its effectiveness. The general idea, then, is to provide a spontaneous form of local community policing in line with more traditional approaches such as 'neighbourhood watch' initiatives.

The participants in online crowdsourced investigations may deny any intention of trespassing on law enforcement prerogatives, or they can be presented as auxiliaries whose powers are clearly delimited by the authorities, but the simple fact of engaging in such activities draws them into a repressive logic. Identifying someone as suspect is an essential stage of an investigation, but doing so *publicly* necessarily involves eluding the many barriers which institutions have set up to avoid harming innocents and also to separate establishment of the facts from any decision on the appropriate sanction (if any). Discussions in threads and comments of online investigations often reveal that after a while hesitations and objections are raised, which drive the initiators or the most involved to justify their action and sometimes set up limits or even rules to prevent excesses. The collectives formed around these various issues therefore always seem to be rediscovering the relevance of formal guidelines and procedures. For instance, in the case of the *Reddit Bureau of Investigation*, research has shown that participants tend to see their work as a complement to law-enforcement, and distinct from vigilantism because violence and self-justice are frowned upon[43]. However, to avoid harassment, defamation and false accusations the forum has set up rules banning 'witch hunts' and the sharing of personal information about suspects.

The motives of amateur investigators may vary but, in any case, their subjectivity is not held in check by formal procedures and may easily slide into self-righteous expectations of justice, with few barriers against violations of privacy and the expression of prejudices such as racial discrimination. Attempts to solve a given case can easily be conflated with the expression of a collective moral judgement conducive to unwarranted social control.

Conversely, institutions and procedures for law enforcement are expected to make sure that these two aspects are clearly separated, with on the one hand the investigation and on the other the assessment of its legal consequences, bearing in mind that both the suspect and the informants may be held accountable – the first may face arrest and lawsuits for his deeds, but the second may also face charges of defamation, libel or false testimony.

Hounding

Hounding goes further still and represents perhaps the epitomy of digital vigilantism. Not only does it combine an investigative dimension with a punitive intention, but it also involves a more sustained mobilisation against a specific target, triggered by intense outrage. It is more squarely associated with *naming and shaming* and involves sharing personal information about someone accused of wrongdoing, in order to punish them by presenting them in a negative light. The aim here is no longer to denounce a behaviour, to solve a puzzle or identify a suspect, but to accuse a person publicly and discredit or humiliate them by providing incriminating evidence. In its more extreme manifestations, it hinges on existing practices of online bullying, harassment and sometimes digital sabotage which, however, are here wielded in the name of justice, order or safety (rather than individual malice or revenge).

One of its main tactics known as *doxing*[44] consists in deliberately seeking personal information (sometimes through unauthorized access into an information system) such as phone numbers, addresses, photos, or social security details in order to spread them online. In many cases, with the target clearly identified and made more vulnerable through the release of such information, it can then take the form of a collective chase or hunt, with repeated forms of harassment carried out as a group, leading to intense exposure to insults, humiliations and threats. It may also lead to more organised practices known as *raiding* or *swarming*, which involve pre-existing communities or networks of potential fellow-vigilante, a greater capacity to coordinate and rehearsed types of actions which include massive negative reviews of the target's business, unsolicited pizza deliveries, and 'defacement' or disabling of websites or online services.

In many cases, hounding can be a consequence of the investigative intentions detailed above: solving the offense can easily slip towards a threat to expose the person involved and shame her. In one example, a Facebook user first shared the picture of her son's bike on December 9, 2015, along with the status: 'STOLEN!!!!! This bike was stolen from behind KFC in [redacted] around 9 this morning any info please contact myself or [redacted]. Police have been informed. Small reward for return. SHARE SHARE SHARE make this bike too

hot [to] handle. Thanks.' Shortly afterwards the user wrote in the comments: 'Thanks 2 the power of social media we now know who has my sons bike the only problem now is finding him'. In a second post on December 11, 2015 (figure 5), the user shared the picture of the bike again, this time with a more threatening status: 'To the person that has got my son's bike bring it back or I will b going 2 the police in the morning with your name pic and message. Plus I will name and shame u on here. Please everybody SHARE SHARE SHARE. Love the power of social media'. The post was shared over 15,000 times in about 10 days.

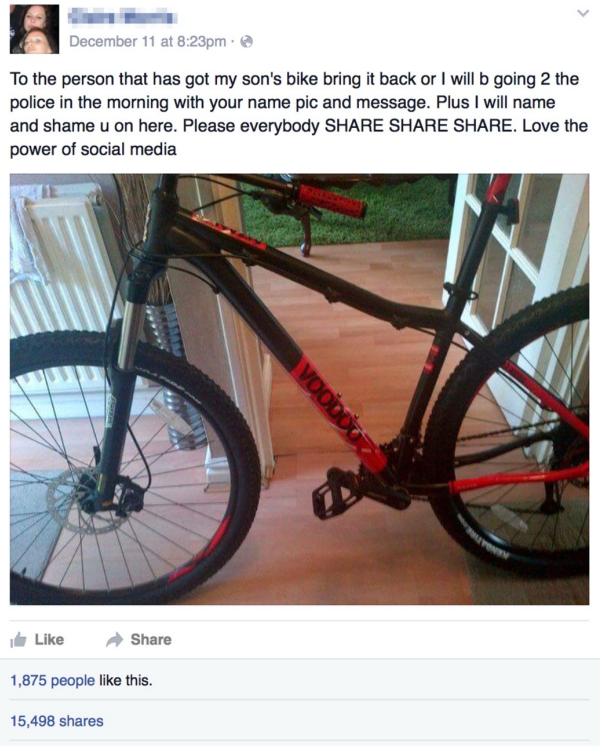

Figure 5. Facebook publication, captured 28 December 2015.

Although the original aim was to find information about a potential thief, once a suspect was identified the platform was used to exert two types of threat: first that evidence would be passed on to the police, and second that another step would be taken with the suspect being publicly shamed – in this case, mainly within the local community he belongs to. Since the bike wasn't returned, a third post on the next day displayed a picture of the suspect along with a name and messages collected from the Facebook messaging functionality.

In other instances, hounding is not the result of an investigation, but consists in a deliberate attempt to retaliate against someone – either because legal channels are uncertain or in order to inflict a more direct form of punishment. In the following case a builder was accused by his client of repeatedly delaying the work he had been paid for. After the client

and his family vented criticism on Facebook, the builder accused them of 'defaming' him. Eventually the client filed a formal complaint with the police, and his daughter not only denounced the builder but also provided complete evidence of his alleged wrongdoings: his name, picture and business address were shared, along with the correspondence and especially text messages detailing unanswered queries and repeated excuses for the delays (figure 6). As much information as possible was published in order to back the accusers' claims and strengthen her case.

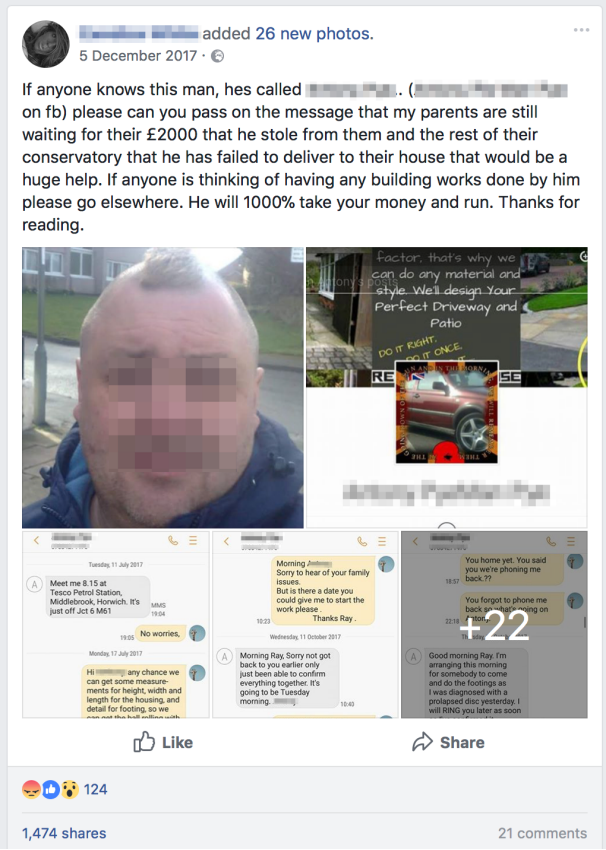

Figure 6. Facebook publication, captured 23 March 2018.

The number of shares can be an indication of the extent to which the person has been shamed, although its effect may be dramatically increased if the person involved is already a public figure, if the case is picked up by someone with a large social media following, and/or if mainstream media decide to cover it. In some cases it may change the course of an individual trajectory: for instance when a dashcam video of Uber CEO Travis Kalanick mistreating an Uber driver was published online, thus playing an important part in precipitating his demise after an already long string of damaging events and revelations[45].

Most of the time, such denunciations remain limited in scope, and are perhaps the most closely related to Brunton's parallel with earlier forms of *charivari* mentioned above. By

calling the crowd to bear witness to the injustice and seeking a collective judgement and backing, such situations normally prompt only limited forms of public chastisement. However, depending on the degree of outrage, the scale of public exposure and the determination to carry out coordinated retributive actions, they may also take greater proportions.

A number of situations involve perceived offences or crimes which have been caught on film and shared online – either deliberately by the authors themselves (unaware that this might be damaging for them), or by third parties who have decided to publicize what they witnessed. The visual evidence sparks outrage and leads viewers to track down the suspects not only to identify them, but also with a clear intention of ensuring that they are punished – either by legal means or through intense shaming and sometimes direct coercion. Common incidents of this type are cases of animal abuse, such as when CCTV footage of a woman dumping a cat in a dustbin in Coventry, UK, was shared on YouTube, totalling over 1,5 million views, leading to an arrest and a modest fine but enduring humiliation and even death threats ('Cat bin woman', 2010). Likewise, a young man was filmed – willingly this time – throwing a kitten against a wall in Marseille, France, igniting public outcry, petitioning and appeal to the police via Twitter, also leading to the identification and arrest of a suspect, and an unusually strong prison conviction ('*Oscar le chaton*', 2014). A famous case involved the killing of a lion from Hwange National Park in Zimbabwe, by an American dentist and recreational hunter who wasn't officially charged, but was publicly condemned, received massive amounts of hate messages, had his personal and business details disclosed, his dental practice negatively rated and website taken down, and his home vandalised ('Cecil the lion', 2015). In China a well-known case saw a young woman filmed crushing kittens with her stilettos ('Hangzhou Kitten Killer', 2006), leading to a form of witch hunt often referred to as 'human flesh search engines', sparked by cases of animal cruelty but also as a means of hunting down hit-and-run drivers or corrupt officials[46]. As well as hitting a particularly sensitive chord in public opinion and drawing exceptional media attention, cases of animal cruelty can also often rely on existing networks of animal rights activists to spread the information and amplify the mobilisation.

More radical cases of hounding involve the targeting of a category of individuals labelled as 'deviant'. Child sex offenders and abusers, for instance, are particularly prone to stirring profound emotions of moral outrage and public mobilisation[47], and are often presented as a form of exceptional crime justifying extra-judicial forms of punishment such as naming and shaming in the press (e.g. by the tabloid *News of the World* in the UK in 2000)

or appeals to death penalty. The internet has encouraged forms of 'paedophile hunting'[48], which can take the form of 'online civilian policing groups' seeking to collaborate with law enforcement[49]: volunteers may scour social media to detect suspicious behaviour or identify registered sex offenders, and sometimes operate fake online profiles and pose as children on chat websites in order to trap child groomers and suspected paedophiles. Any evidence collected during these sting operations such as chat logs, screen captures or posted contents may then be passed on to the police, raising controversy due to the difficulty of using such information in court (especially insofar as entrapment might be considered an incitement to commit an offence and abuse of process), the possible interference with existing police operations, and more generally the mistakes, unintended consequences and general risks which these types of 'moral crusades' can involve. In some contexts, the hunters may also attempt to punish the suspects themselves, for instance by disclosing more widely the collected evidence and leveraging social media to inflict public shaming; sometimes the suspected paedophiles may even be beaten up or coerced into recording humiliating confessions on video, as has been observed in Russia[50].

One further example involves 'scambaiting'. This consists in responding to pervasive online scams such as 'advance fee frauds' (emails which attempt to trick the receiver into believing they will receive large transfers of money, for which they should however pay an advance fee) by outmanoeuvring their authors, engaging in lengthy interactions intending to frustrate and ridicule them, and finally tricking them into revealing personal information. The so-called 'Nigerian scam' or '419 scam', named after the section of the Nigerian Criminal Code addressing these types of frauds (and although Nigerians now make up only a small proportion of perpetrators), is a major variant of such schemes[51]. According to Dara Byrne it has generated particularly intense scambaiting activity since the late 1990s[52], and is usually presented as a 'fun' form of community service and as a way to deter such types of cross-border criminal activity, against which official law enforcement is considered helpless.

However, one of the main ways of exerting 'justice' is to get the scammers to send a copy of their ID or of a humiliating picture of themselves, often holding a prop or a sign with a demeaning inscription (generally of a sexual nature), in some cases appearing nude or even fulfilling degrading and physically painful requests (such as having the baiter's pseudonym tattooed on their body in order to scar and mark it), and then publishing them online in a 'hall of shame' and brandishing them as 'trophies'[53]. Having analysed these practices across time, Byrne notes that '419 digilante tactics bear comparison to the rich history of anti-black vigilantism, particularly in an American context' and draws parallels with

the ways in which vigilante committees established in the United States from the mid-1700s onwards eventually turned into lynching directed towards minorities – and particularly African-Americans. Beyond the initial claim to fend off crime and ensure security, lynching 'evolved into a specifically racialized form of violence' which 'served as a popular form of cultural entertainment for the communities that practiced them' and primarily relied on 'visually sensational aspects' to assert social order as racial hierarchy. Similarly, on websites such as 419 Eater, tens of thousands of registered users post on the forums section and discuss their feats: 'Like lynching memorabilia, the rhetorical power of these trophies is their socializing function since they are critical in establishing the code of ethics and the reward system for the community.'[54] For Lisa Nakamura such practices are meant to 'police the digital primitive' by leveraging memetic culture and its properties for building group identity[55]. They can therefore aptly be equated with *digital lynching* as a possible development inherent in digital vigilantism.

Some of the most well-known cases of hounding, however, may be associated with the loosely defined Anonymous movement. It was particularly active from 2008 to 2012 and attracted considerable media attention as a controversial figure of 'hacktivism'[56], engaging in online direct actions in order to defend a number of 'causes'. The movement is hard to classify due to the great diversity of targets (often taken up on a voluntary basis by different sub-groups), its equivocal communication and collective identity, and the ambivalence of its repertoires of action. Indeed, the latter can sometimes be understood as forms of *protest and dissent* in line with more traditional practices of media activism, however in many cases they also constitute an attempt at effecting *direct punishment* of individuals or institutions through various means, involving either denunciation or disruption: disclosing personal information, sometimes after hacking into information systems (doxing); altering of websites ('defacement'); obstruction of websites (e.g. by launching distributed denial-of-service or DDoS attacks); or otherwise converging *en masse* towards a target and flooding it with negative comments or offensive content.

The movement originated on the 4chan image board, known for its tolerance towards the most controversial material and opinions, and where users are all logged in as 'Anonymous' by default. 4chan users, meeting and socializing on the forum and often coordinating via *Internet Relay Chat* (IRC), have been known to carry out different forms of juvenile pranks and hoaxes: flooding forums or comments sections on social media with derogatory messages or disturbing content, crashing or defacing websites, or having quantities of pizzas or taxis sent to a target[57]. In some circumstances such 'trolling' may take

a vicious turn, and has been characterized by Gabriella Coleman as 'the targeting of people and organizations, the desecration of reputations, and the spreading of humiliating information'[58].

Whilst they were initially carried out as a form of malicious humour ('for the lulz'), after 2008 some of the same methods were used against more political targets such as the Church of Scientology, in what Coleman describes as a coming of age of Anonymous. Indeed, not only did the movement protest through defacing and taking down websites, prank calls and Google bombing, or doxing of senior members of the Church, but for the first time it also organised carnivalesque physical demonstrations throughout the world, and established a more explicit common cause through the (loosely defined) defence of free speech. In this occasion its signature motto, both threatening and impish, was forged: 'We are Anonymous. We are legion. We do not forgive. We do not forget. Expect us.' The famous Guy Fawkes mask from the *V for Vendetta* graphic novel was also used for the first time – initially to prevent identification by Scientology. Such symbolic elements eventually played a key role in establishing the collective banner of a purported multitude, vengeful yet bestowed with a form of popular legitimacy. The group also adopted increasingly elaborate communication techniques – both externally in terms of public relations (by publishing videos, messages on social media, and 'press releases' on pastebins), and internally through sustained covert interactions (socialisation on forums, setting up of websites as resources, and coordination via IRC channels).

From then on, many other campaigns – known as 'Ops' – were carried out, of varying scale and against diverse targets[59]. One of the most prominent in 2010-2011 sought to retaliate against antipiracy organisations and corporations though multiple DDoS attacks ('Operation Payback') before mutating into support of WikiLeaks ('Operation Avenge Assange'). Other campaigns involved, for instance, taking down child pornography servers and websites, releasing the names of hundreds of users and inviting the FBI to take action ('Operation DarkNet', 2011). Major operations were also directed against the suspected authors of gang rapes in Steubenville, Ohio and in Maryville, Missouri, whose identity along with incriminating evidence where made public, along with alleged dysfunctions and cover ups by the local authorities, generating heated debate ('Op Roll Red Roll' and 'Op Maryville', 2013)[60]. In many cases however, interventions by Anonymous ultimately proved ineffective, triggered unintended consequences (such as overexposing a victim), or purely and simply doxed the wrong person.

Due to their coordinated nature, potential violence and more marked incursions into illegal territories, hounding actions generate greater controversy and usually require more articulate forms of justification. They are therefore usually tied to a 'cause' (e.g. animal rights) which participants will claim to defend, or will explicitly appeal to the notion that 'justice should be done' (although in the case of Anonymous, a large part of its members resisted following the 'moral turn' of the movement and were primarily interested in wreaking havoc). While institutions (such as Scientology) may sometimes come under fire, in which case the action may involve a political dimension, hounding is characterized by pronounced and sustained targeting of individuals or categories of individuals (e.g. 'deviants') who will be named, shamed, and often harassed as a consequence – a chain of events which may be substantially amplified by the echo chamber provided by mainstream media. In some cases, authorities may be unable to ignore the public attention and hounding may encourage police intervention and legal developments. In other cases, the state's prerogatives (justice and policing) are challenged to the point that participants will take greater care to remain anonymous and law enforcement will attempt to reassert its power (e.g. by unmasking Anonymous members and pressing charges against them).

Organised leaking

Finally, *organised leaking* is primarily directed at *institutions or organisations*, and involves a higher degree of structuration through the setting up specific processes and technological tools intended to encourage and manage the documenting of problematic situations or the disclosure of confidential – and potentially incriminating – information. It is thus also different from one-off whistleblowing initiatives on the part of 'insiders' witness to unethical practices, and can be understood as *architectures of denunciation*, which usually include protecting sources through anonymisation and favouring the widespread publication of this information, but do not necessarily involve a high degree of technical sophistication. Efforts are longer term however, and contrary to other forms of digital vigilantism, the justifications for such practices are usually grounded in ethical considerations with a high degree of generality, since 'systemic' issues tend to be targeted rather than individuals (although this may involve leveraging individual apprehensions of shame or discredit). Indeed, beyond the denunciation of specific situations, such enterprises aim at challenging the overall structure of existing institutions responsible for ensuring, say, public accountability (e.g. journalism) or scientific integrity (e.g. peer review publishing), and to correct their shortcomings.

The archetype for this kind of activity is the WikiLeaks organisation, which devised and set up a system geared towards facilitating and securing the anonymous disclosure of vast amounts of information – such as the US military reports of the wars in Iraq and Afghanistan as well as diplomatic cables, all published in 2010. Despite having now come under intense criticism for promoting various conspiracy theories, providing misleading information and for helping to derail the 2016 American presidential election, WikiLeaks was initially presented as expanding the modern liberal project of publicity and as a crucial evolution of journalism[61], a new form of civil disobedience[62] – or a novel instance of vigilantism[63]. WikiLeaks itself has always sought to present its activities as driven by the idea that radical transparency could help to prevent abusive or otherwise unethical behaviour on the part of institutions such as governments or corporations – thus deserving the status (and legal protection) of journalism. Although it has now become clear how such anonymous information disclosures can be manipulated or be otherwise ineffective or damaging, WikiLeaks has long served as a model for promoting a form of self-regulation driven by the fear of leaks.

Other projects have been devised in order to facilitate the disclosure of dishonest or otherwise problematic behaviour in more specific fields. This is the case for instance in science with approaches in terms of 'post-publication peer review', which have led to the setting up of websites such as *Retraction Watch* and *PubPeer* which encourage the anonymous reviewing of already published articles (although *PubPeer* is moderated and accusations of fraud are suppressed when insufficient evidence is provided), thus leveraging the potential discredit associated with scientific misconduct: 'We have seen a new pack of watchdogs coming to the fore. Most of them are internet-based, are fed by grass-root researchers, and employ a variety of different shaming techniques. (…) This pack is here to stay, and it foreshadows one future of scientific evaluation.'[64] Papers may be criticized for presenting forgeries (especially image manipulation in biology) or low-quality work, and on *PubPeer* authors will automatically be invited to respond to the criticism – sometimes leading to thorough argumentation but also heated discussions. *PubPeer* also provides a plugin allowing users to receive alerts, if an article has been commented, when browsing journal websites – thus ensuring that the shaming process or at least the attention generated by a paper will be carried over to its source. Many articles have already been retracted and some prestigious academics have seen their reputation shattered, while others have sued the owners of the website for defamation. Critics have been vocal against the prospect of a 'vigilante science'[65] and the unaccountability of anonymous criticism, while its proponents have defended their

position as 'vigilant scientists'[66], arguing that: 'platforms like *PubPeer* can help ensure that cheating, once discovered, has lasting consequences, tilting the balance of benefits towards honest, high-quality research.'[67]

General overview

These four ideal-typical constructs present different degrees and forms of digital vigilantism associated with various repertoires of online direct action, as summarised in table 1. However, as has also been shown it is quite frequent that situations either straddle different categories or slip from one to another – for instance flagging may result in targeting individual persons who then become hounded, and investigating someone may also lead to hounding her. Moreover, both flagging and organised leaking may be understood as more peripheral types, insofar as they normally don't involve the retributive targeting of individuals, while the core of digital vigilantism is represented by investigating and hounding. As mentioned from the outset, this typology is primarily intended as a heuristic device, which stresses the inherent logic driving these different types of interactions in order to provide an overall picture of digital vigilantism.

Conclusion

The justifications for engaging in digital vigilantism are varied and range from norms of civility to crime control, expectations of order and quests for justice: as with traditional forms of vigilantism, they can never be simply a matter of personal revenge, and must involve a form or other of generalization. However, the specificities of the digital environment have redefined 'speech acts', insofar as merely voicing a concern online – and documenting it in various manners – may already be a step towards redress.

Indeed, public denunciations – and more generally, situations where the distinction between private and public suddenly collapses – have always involved a dimension of potential violence, symbolic (through disgrace and dishonour) and sometimes real (through harassment and even physical coercion borne of resentment)[68]. In a digital setting, this line may be more easily trodden. Especially in cases pertaining to the two core types of digital vigilantism, investigating and hounding, mere publication can in itself entail punitive consequences for suspects, by concentrating negative publicity long before the complete facts are established. This is compounded when situations are relayed by mainstream media and journalists who, simply by attempting to document an event, in fact widely expand its

audience and thus increase the shame and humiliation for the targets – often based on very little information and with no proper understanding of context[69].

Obviously, as is often already the case when 'trial by media' is wielded, key legal notions of presumption of innocence are subverted in the process, along with their many implications – burden of proof resting with accusers, right of confrontation and right of counsel for the accused etc. Digital vigilantism is easily set off but is rife with mistakes, unintended consequences, self-righteousness, collateral damages, interference with law enforcement, and sometimes outright violence[70]. Despite this lack of accountability, in the context of generalized mutual surveillance and the drive to self-regulation mentioned at the outset, it is tempting to engage in it – and thus attempt to bypass or compete with institutional channels for law enforcement, uncovering of facts and establishment of justice.

Indeed, understood as both a moral and a political phenomenon, digital vigilantism is an increasingly accessible means of converting outrage, security concerns or assumptions of injustice into effective action online. As an informal but potentially powerful mode of social control, it upsets institutions such as journalism, legal systems and policing which are normally entrusted with revealing, judging and punishing transgressive behaviours in a democratic context. It represents both a challenge to these institutions – and in some circumstances an answer to their shortcomings. Digital vigilantism comes in many different shades, in terms of its concrete manifestations and intensity, and attention should always be paid to the contextual specificities of its occurrence. The typology presented here provides an account of the main courses of action involved when seeking to resolve situations of conflict directly in a digital environment, as well as their underlying logic. It is hoped that it will serve as a tool for further research and for helping to assess its moral implications, legal consequences and democratic compatibility – or lack thereof.

Notes

[1] See for instance Hague and Loader eds., *Digital Democracy*; Shane ed., *Democracy Online*; Dahlberg and Sapiera eds., *Radical Democracy and the Internet*; and Bennett ed., *Civic Life Online*.

[2] Castells, *Networks of Outrage and Hope*.

[3] Levmore and Nussbaum, *The Offensive Internet*; and Citron, *Hate Crimes in Cyberspace*.

[4] I would like to thank the two anonymous reviewers for their careful reading of the manuscript and their valuable suggestions and comments.

[5] Moncada, 'Varieties of Vigilantism'.

[6] Rosenbaum and Sederberg, *Vigilante Politics*.

[7] Kirkpatrick, *Uncivil Disobedience*.

[8] Johnston, 'What Is Vigilantism?'

[9] Trottier, 'Digital Vigilantism As Weaponisation of Visibility.'

[10] Brunton, *Spam*, 46-47.

[11] Andrejevic, 'The Work of Watching One Another'; and Albrechtslund, 'Online Social Networking As Participatory Surveillance.'

[12] Marwick and boyd, 'I Tweet Honestly, I Tweet Passionately.'

[13] Auray, 'Online Communities and Governance Mechanisms.'

[14] McLaughlin, Osborne and Smith, 'Standards of Conduct on Usenet.'

[15] Cardon, 'Discipline but Not Punish: The Governance of Wikipedia.'

[16] Crawford and Gillespie, 'What Is a Flag For?'

[17] Loveluck, 'The Internet: A Society Against the State?'

[18] Less than half of them could be mentioned in this article for lack of space.

[19] Boellstorff, Nardi and Pearce, *Ethnography and Virtual Worlds*; Markham and Baym eds. *Internet Inquiry*; and Hine, 'Virtual Ethnography: Modes, Varieties, Affordances.'

[20] In particular, public Facebook groups and pages were scraped using the Netvizz application developed by Bernhard Rieder and based on the Facebook API, which allowed the collection of timestamped posts and comments, along with reaction, sharing and comment metrics (see Rieder, 'Studying Facebook Via Data Extraction').

[21] Weber classically defines ideal types as "formed by the one-sided *accentuation* of one or more points of view and by the synthesis of a great many diffuse, discrete, more or less present and occasionally absent *concrete individual* phenomena, which are arranged according to those one-sidedly emphasized viewpoints into a unified *analytical* construct." (Weber, '"Objectivity" in social science and social policy', 90).

[22] Boltanski and Thévenot, *On Justification*; and Boltanski, *Love and Justice As Competences*.

[23] De Blic and Lemieux, 'Le scandale comme épreuve'.

[24] Tilly, *The Contentious French*; Tilly, *The Politics of Collective Violence*; and Tarrow, *Power in Movement*.

[25] Tilly, *The Contentious French*; McAdam, Tarrow and Tilly, *Dynamics of Contention*, Ch. 5.

[26] Norman, *The Design of Everyday Things*.

[27] Van Laer and Van Aelst, 'Internet and Social Movement Action Repertoires.'

[28] Sargsian, 'The Dash Cam Phenomenon'.

[29] One such masked and anonymous cyclist called 'Cinquante Euros' has attracted media attention by setting up a YouTube channel (https://www.youtube.com/channel/UC8rE-HbmlCjj_xWwPm13grQ, over 13,000 subscribers), along with Facebook and Twitter accounts, where such incidents are documented.

[30] On July 4, 2018, https://www.facebook.com/PassengerShaming totalled over 500,000 'likes' and https://www.instagram.com/passengershaming had nearly 670,000 subscribers.

[31] https://www.facebook.com/pg/Etaples/about/

[32] 'Étaples : ils dénoncent des incivilités sur Facebook sans se dévoiler', *La Voix du nord*, April 20, 2016.

[33] '"Deal Safari" à Bordeaux : première condamnation', *Le Monde.fr*, May 4, 2013.

[34] Yardley et al., 'What's the Deal with 'websleuthing'?', 82.

[35] https://www.reddit.com/r/RBI/ and https://www.reddit.com/r/UnresolvedMysteries/ (over 68,000 and 446,000 subscribers respectively as of July 2018).

[36] Yardley et al., 'What's the Deal with 'websleuthing'?', 97-102.

[37] Nhan, Huey and Broll, 'Digilantism'.

[38] Lally, 'Crowdsourced Surveillance and Networked Data.'

[39] 'Reflections on the recent Boston crisis', *Reddit.com* blog, April 22, 2013. https://redditblog.com/2013/04/22/reflections-on-the-recent-boston-crisis/

[40] Massanari, *Participatory Culture, Community, and Play*.

[41] Schneider and Trottier, 'The 2011 Vancouver Riot and the Role of Facebook in Crowd-sourced Policing.'

[42] Schneider and Trottier, 'The 2011 Vancouver Riot and the Role of Facebook in Crowd-sourced Policing', 68.

[43] Myles, Millerand, and Benoit-Barné. 'Résoudre des crimes en ligne.'

[44] Douglas, 'Doxing.'

[45] Newcomer, E. and Stone, B. 'The Fall of Travis Kalanick Was a Lot Weirder and Darker Than You Thought.' *Bloomberg Businessweek*, January 18, 2018.

[46] Herold, 'Development of a Civic Society Online?'; and Gao and Stanyer, 'Hunting Corrupt Officials Online.'

[47] Meyer, *The Child at Risk*.

[48] Campbell, 'Policing paedophilia.'

[49] Huey et al., ''Uppity civilians' and 'cyber-vigilantes''.

[50] Favarel-Guarrigues, 'Justiciers amateurs et croisades morales en Russie contemporaine.'

[51] Brunton, *Spam*, 101-110.

[52] Byrne, '419 digilantes and the frontier of radical justice online', 71-72.

[53] See for instance https://www.419eater.com/html/trophy_room.htm and https://www.419eater.com/html/hall_of_shame.htm.

[54] Byrne, '419 digilantes and the frontier of radical justice online', 77.

[55] Nakamura, ''I WILL DO EVERYthing that am asked': scambaiting, digital show-space, and the racial violence of social media'.

[56] Klein, 'Vigilante media.' On hacktivism see Jordan, *Activism!*

[57] Phillips, *This Is Why We Can't Have Nice Things*.

[58] Coleman, *Hacker, Hoaxer, Whistleblower, Spy*, 19.

[59] The following Wikipedia page presents a detailed though incomplete list of over a hundred actions undertaken from 2006 to 2017: https://en.wikipedia.org/wiki/Timeline_of_events_associated_with_Anonymous

[60] See Christensen, 'All politics is local.'

[61] Benkler, 'WikiLeaks and the Networked Fourth Estate.'

[62] Züger, 'Re-thinking Civil Disobedience.'

[63] Heemsbergen, 'Designing Hues of Transparency and Democracy After WikiLeaks.'

[64] Didier and Guaspare-Cartron, 'The New Watchdogs' Vision of Science.'

[65] Blatt, 'Vigilante Science.'

[66] *PubPeer*, 'Vigilant Scientists.'

[67] *PubPeer*, 'A Crisis of Trust.'

[68] Boltanski and Claverie, 'Du monde social en tant que scène d'un procès.'

[69] McBride, 'Journalism and public shaming.'

[70] Chang et al., 'Citizen Co-Production of Cyber Security'.

Bibliography

Albrechtslund, A. 'Online Social Networking As Participatory Surveillance.' *First Monday [online]* 13, no. 3 (2008).

Andrejevic, M. 'The Work of Watching One Another: Lateral Surveillance, Risk, and Governance.' *Surveillance & Society* 2, no. 4 (2005): 479-497.

Auray, N. 'Online Communities and Governance Mechanisms.' In *Governance, Regulation and Powers on the Internet*, edited by E. Brousseau, M. Marzouki and C. Méadel, 211-231. Cambridge and New York: Cambridge University Press, 2012.

Benkler, Y. 'WikiLeaks and the Networked Fourth Estate.' In *Beyond WikiLeaks. Implications for the Future of Communications, Journalism and Society*, edited by B. Brevini, A. Hintz and P. McCurdy, 11-34. Basingstoke and New York: Palgrave Macmillan, 2013.

Bennett, W.L., ed. *Civic Life Online. Learning How Digital Media Can Engage Youth*. Cambridge, MA and London: MIT Press, 2008.

Blatt, M. R. 'Vigilante Science.' *Plant Physiology* 169, no. 2 (2015): 907-909.

Boellstorff, T., B. A. Nardi, and C. Pearce. *Ethnography and Virtual Worlds. A Handbook of Method*. Princeton, NJ and Oxford: Princeton University Press, 2012.

Boltanski, L., and L. Thévenot. *On Justification. Economies of Worth*. Princeton, NJ and Oxford: Princeton University Press, 2006 [1991].

Boltanski, L. *Love and Justice As Competences. Three Essays on the Sociology of Action*. Cambridge and Malden, MA: Polity, 2012 [1990].

Boltanski, L., and E. Claverie. 'Du monde social en tant que scène d'un procès.' In *Affaires, Scandales Et Grandes Causes. De Socrate À Pinochet,* edited by L. Boltanski, E. Claverie, N. Offenstadt and S. Van Damme. Paris: Stock, 2007.

Brunton, F. *Spam. A Shadow History of the Internet*. Cambridge, MA and London: MIT Press, 2013.

Byrne, D. N. '419 Digilantes and the Frontier of Radical Justice Online.' *Radical History Review* no. 117 (2013): 70-82.

Campbell, E. 'Policing Paedophilia: Assembling Bodies, Spaces and Things.' *Crime, Media, Culture* 12, no. 3 (2016): 345-365.

Castells, M. *Networks of Outrage and Hope. Social Movements in the Internet Age*. Cambridge and Malden, MA: Polity Press, 2012.

Cardon, D. 'Discipline but Not Punish: The Governance of Wikipedia.' In *Normative Experience in Internet Politics,* edited by F. Massit-Folléa, C. Méadel and L. Monnoyer-Smith, 211-232. Paris: Transvalor/Presses des Mines, 2012.

Chang, L.Y.C, Zhong, L.Y., and Grabosky, P.N. 'Citizen Co-Production of Cyber Security: Self-Help, Vigilantes, and Cybercrime.' *Regulation & Governance* 12 (2018): 101-114.

Christensen, C. 'All Politics Is Local. Anonymous and the Steubenville/Maryville Rape Cases.' In *The Routledge Companion to Social Media and Politics*, edited by A. Bruns, G. Enli, E. Skogerbø, A. O. Larsson and C. Christensen, 153-164. London and New York: Routledge, 2016.

Citron, D. K. *Hate Crimes in Cyberspace*. Cambridge, MA and London: Harvard University Press, 2014.

Coleman, G. *Hacker, Hoaxer, Whistleblower, Spy. The Story of Anonymous*. London and New York: Verso, 2014.

Crawford, K., and T. Gillespie. 'What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint.' *New Media & Society* 18, no. 3 (2016): 410-428.

Dahlberg, L., and E. Siapera, eds. *Radical Democracy and the Internet. Interrogating Theory and Practice*. Basingstoke and New York: Palgrave Macmillan, 2007.

De Blic, D., and C. Lemieux. 'Le scandale comme épreuve. Éléments de sociologie pragmatique.' *Politix* 18, no. 71 (2005): 9-38.

Didier, E., and C. Guaspare-Cartron. 'The New Watchdogs' Vision of Science: A Roundtable with Ivan Oransky (*Retraction Watch)* and Brandon Stell (*PubPeer*).' *Social Studies of Science* 48, no. 1 (2018): 165-167.

Douglas, D. M. 'Doxing: A Conceptual Analysis.' *Ethics and Information Technology* 18, no. 3 (2016): 199-210.

Favarel-Guarrigues, 'Justiciers amateurs et croisades morales en Russie contemporaine.' *Revue française de science politique* 68, no. 4 (2018): 651-667.

Gao, L., and J. Stanyer. 'Hunting Corrupt Officials Online: The Human Flesh Search Engine and the Search for Justice in China.' *Information, Communication & Society* 17, no. 7 (2014): 814-829.

Hague, B.N., and B. Loader, eds. *Digital Democracy. Discourse and Decision Making in the Information Age*. London and New York: Routledge, 1999.

Heemsbergen, L. J. 'Designing Hues of Transparency and Democracy After WikiLeaks: Vigilance to Vigilantes and Back Again.' *New Media & Society* 17, no. 8 (2015): 1340-1357.

Herold, D. 'Development of a Civic Society Online? Internet Vigilantism and State Control in Chinese Cyberspace.' *Asia Journal of Global Studies* 2, no. 1 (2008): 26-37.

Hine, C. 'Virtual Ethnography: Modes, Varieties, Affordances.' In *The SAGE Handbook of Online Research Methods*, edited by N. G. Fielding, R. M. Lee and G. Blank, 257-270. London, Thousand Oaks, CA and New Delhi: Sage, 2008.

Huey, L., Nhan, J., and Broll, R. "Uppity Civilians' and 'Cyber-Vigilantes': The Role of the General Public in Policing Cyber-Crime.' *Criminology & Criminal Justice* 13 no. 1 (2013): 81-97.

Johnston, L. 'What Is Vigilantism?' *British Journal of Criminology* 36, no. 2 (1996): 220-236.

Jordan, T. *Activism! Direct Action, Hacktivism and the Future of Society*. London: Reaktion Books, 2001.

Klein, A. G. 'Vigilante Media: Unveiling Anonymous and the Hacktivist Persona in the Global Press.' *Communication Monographs* 82, no. 3 (2015): 379-401.

Kirkpatrick, J. *Uncivil Disobedience. Studies in Violence and Democratic Politics*. Princeton, NJ: Princeton University Press, 2008.

Lally, N. 'Crowdsourced Surveillance and Networked Data.' *Security Dialogue* 48, no. 1 (2017): 63-77.

Levmore, S., and M.C. Nussbaum, eds. *The Offensive Internet. Privacy, Speech, and Reputation*. Cambridge, MA: Harvard University Press, 2011.

Loveluck, B. 'The Internet: A Society Against the State? Informational Liberalism and Political Economies of Self-organization in the Digital Regime.' *Réseaux*, no. 192 (2015): II-XXXIV.

Markham, A. N. and N. K. Baym, eds. *Internet Inquiry. Conversations about Method*. Los Angeles, CA and London, Sage, 2009.

Marwick, A. E., and d. boyd. 'I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience.' *New Media & Society* 13, no. 1 (2011): 114-133.

Massanari, A. L. *Participatory Culture, Community, and Play. Learning From Reddit*. New York: Peter Lang, 2015.

McAdam, D., Tarrow S. and Tilly C. *Dynamics of Contention*. Cambridge and New York: Cambridge University Press, 2001.

McBride, K. 'Journalism and Public Shaming: Some Guidelines.' *Poynter*, March 11, 2015. https://www.poynter.org/news/journalism-and-public-shaming-some-guidelines

McLaughlin, M. L., K. K. Osborne, and C. B. Smith. 'Standards of Conduct on Usenet.' In *Cybersociety. Computer-Mediated Communication and Community*, edited by S. G. Jones, 90-111. London, Thousand Oaks, CA and New Delhi: Sage, 1995.

Meyer, A. *The Child at Risk. Paedophiles, Media Responses and Public Opinion*. Manchester: Manchester University Press, 2007.

Moncada, E. 'Varieties of Vigilantism: Conceptual Discord, Meaning and Strategies.' *Global Crime* 18, no. 4 (2017): 403-423.

Myles, D., F. Millerand, and C. Benoit-Barné. 'Résoudre des crimes en ligne. La contribution de citoyens au *Reddit Bureau of Investigation*.' *Réseaux* no. 197-198 (2016): 173-202.

Nakamura, L. "I WILL DO EVERYthing That Am Asked': Scambaiting, Digital Show-Space, and the Racial Violence of Social Media'. *Journal of Visual Culture* 13 no. 3 (2014): 257-274.

Nhan, J., L. Huey, and R. Broll. 'Digilantism: An Analysis of Crowdsourcing and the Boston Marathon Bombings.' *British Journal of Criminology* 57, no. 2 (2017): 341-361.

Norman, D. A. *The Design of Everyday Things*. Original title: *The Psychology of Everyday Things*. London: MIT Press, 1998.

Phillips, W. *This Is Why We Can't Have Nice Things. Mapping the Relationship Between Online Trolling and Mainstream Culture*. Cambridge, MA: MIT Press, 2015.

PubPeer. 'A Crisis of Trust.' July 27, 2014. http://blog.pubpeer.com/?p=164

PubPeer. 'Vigilant Scientists.' October 5, 2015. http://blog.pubpeer.com/?p=200

Rieder, B. 'Studying Facebook Via Data Extraction: The Netvizz Application.' In *Proceedings of the 5th Annual ACM Web Science Conference*. New York: ACM, 2013.

Rosenbaum, H.J. and P.C. Sederberg, eds. *Vigilante Politics*. Philadelphia, PA: University of Pennsylvania Press, 1976.

Sargsian, Z. A. 'The Dash Cam Phenomenon: Technology and the Rule of Law in Russia.' *International Journal of Civil Society Law* 11, no. 1 (2013): 46-50.

Schneider, C. J., and D. Trottier. 'The 2011 Vancouver Riot and the Role of Facebook in Crowd-sourced Policing.' *BC studies*, no. 175 (2012): 57-72.

Shane, P.M., ed. *Democracy Online. The Prospects for Political Renewal Through the Internet*. New York: Routledge, 2004.

Tarrow, S. *Power in Movement. Social Movements and Contentious Politics,* 3rd edition ed. Cambridge: Cambridge University Press, 2011.

Tilly, C. *The Contentious French*. Cambridge, MA and London: Harvard University Press, 1986.

Tilly, C. *The Politics of Collective Violence*. Cambridge and New York: Cambridge University Press, 2003.

Trottier, D. 'Digital Vigilantism As Weaponisation of Visibility.' *Philosophy & Technology* 30, no. 1 (2017): 55-72.

Van Laer, J., and P. Van Aelst. 'Internet and Social Movement Action Repertoires. Opportunities and Limitations.' *Information, Communication & Society* 13, no. 8 (2010): 1146-1171.

Weber, M. '"Objectivity" in Social Science and Social Policy.' In *The Methodology of the Social Sciences*, edited by E.A. Shils and H.A. Finch, 50-112. New York: The Free Press of Glencoe, 1949 [1904].

Yardley, E., A. G. T. Lynes, D. Wilson, and E. Kelly. 'What's the Deal with 'websleuthing'? News Media Representations of Amateur Detectives in Networked Spaces.' *Crime, Media, Culture* 14, no. 1 (2018): 81-109.

Züger, T. 'Re-thinking Civil Disobedience.' *Internet Policy Review* 2, no. 4 (2013).