Ethics Inf Technol (2015) 17:175–188
DOI 10.1007/s10676-015-9366-9

ORIGINAL PAPER

# The use of software tools and autonomous bots against vandalism: eroding Wikipedia's moral order?

Paul B. de Laat[1]

Published online: 2 September 2015

**Abstract** English-language Wikipedia is constantly being plagued by vandalistic contributions on a massive scale. In order to fight them its volunteer contributors deploy an array of software tools and autonomous bots. After an analysis of their functioning and the 'coactivity' in use between humans and bots, this research 'discloses' the moral issues that emerge from the combined patrolling by humans and bots. Administrators provide the stronger tools only to trusted users, thereby creating a new hierarchical layer. Further, surveillance exhibits several troubling features: questionable profiling practices (concerning anonymous users in particular), the use of the controversial measure of reputation (under consideration), 'oversurveillance' where quantity trumps quality, and a prospective loss of the required moral skills whenever bots take over from humans. The most troubling aspect, though, is that Wikipedia has become a Janus-faced institution. One face is the basic platform of MediaWiki software, transparent to all. Its other face is the anti-vandalism system, which, in contrast, is opaque to the average user, in particular as a result of the algorithms and neural networks in use. Finally it is argued that this secrecy impedes a much needed discussion to unfold; a discussion that should focus on a 'rebalancing' of the anti-vandalism system and the development of more ethical information practices towards contributors.

**Keywords** Bots · Disclosive ethics · Profiling · Surveillance · Vandalism · Wikipedia

## Introduction

Communities that thrive on the contributions from their respective 'crowds' have been with us for several decades now. The contents involved may be source code, text, or pictures; the products created may be software, news, reference works, videos, maps, and the like. As argued before (de Laat 2012b, using Dutton 2008), the basic parameters of such open content communities are twofold: on the one hand their technological web design, which may range from sharing (1.0), co-contributing (2.0), to co-creation (3.0); on the other hand their conditions of admission to the collaborative work process, which may be fully open or more restricted (say, only for experts). Some telling examples are Linux, Slashdot, Reddit, NowPublic, Wikipedia, and YouTube.

Open channels of communication invite all kinds of contributions to the collective process. Inevitably, they also solicit contents that are disruptive and damaging to the goals of the community: off-topic, inappropriate, improper, offensive, malicious content and the like. Obviously, this issue is most urgent in those open-content communities that focus on co-contribution or co-creation without any restrictions on entry. Most of such 'open collaboration' projects (a term coined by Forte and Lampe 2013), therefore, have had no choice but to develop systems that detect improper contributions and subsequently eliminate them in appropriate ways. In short, *anti-intrusion systems* have been introduced.

Several varieties can be discerned.[1] A default solution is that professional editors of the site police contributions for improper content, before and/or after publication. Most social news sites and citizen journals are a case in point. As

✉ Paul B. de Laat
p.b.de.laat@cerug.nl

[1] University of Groningen, Groningen, The Netherlands

[1] The next two paragraphs mostly follow de Laat (2012a).

a rule, though, the incoming flow of content is simply too large to be handled in this way alone; therefore the crowds are called to the rescue. Such help may take various forms.

First, contributors can be asked to scout for improper content as well. Any purported transgressions are to be *reported* to the managing editors who will deal with them. If necessary, additional workers are hired for the moderation of reported contents. Facebook and YouTube reportedly outsource this kind of moderation globally to small companies that all together employ thousands of lowly-paid workers. Many of the call centres involved, for example, are located in the Philippines (Chen 2014). Secondly, contributors can be invited to *vote* on new content: usually a plus or a minus. It was, of course, Digg that pioneered this scheme ('digging'). As a result of this voting, contributions rise to public prominence (or visibility) more quickly, or fade away more quickly. Quality gets sorted out. Both reporting and voting are common practice now in social news sites and citizen journals. Thirdly, contributors can be invited to *change* contributions that are deemed inappropriate. Usually, of course, this means deleting the content in question. It is common practice in communities that operate according to a wiki format that allows contributors unrestricted write-access. Cases in point are Citizendium, Wikinews, and Wikipedia. Finally, contributors can be invited to spot improper contributions—but only those who have proved themselves trustworthy are allowed to act accordingly. This *hierarchical* option is common practice in communities of open-source software: after identifying them as genuine bugs, only developers (with write-access) may correct such edits by eliminating or fixing them in the main code tree.

In order to present a unified framework in the above I have been talking indiscriminately about improper and/or disruptive content. Such neutral language of course conceals a lot of *variety* between communities concerning what is to be counted as such. Moreover, impropriety and disruption only exist in the eyes of the particular beholder. These observations immediately raise many associated questions. What are the definitions of impropriety in use? Who effectively decides on what is to count as such? Is it the crowds that have been solicited or only a select few? If the latter, how did the select few obtain their position of power: by appointment from above or by some democratic procedure? Ultimately the crucial question is: do such decisions on what counts as proper content bring us any closer to the particular 'truth' being sought?

Obviously, the answers are bound to vary depending on which type of community is under scrutiny. Let me just state here a few, select generalities. As concerns the kind of content being pursued, some communities are after an exchange of opinions about topical issues (social news, citizen journals), others are on a quest for the 'facts' about certain fields of interest (encyclopedias) or for source code that meets certain technical criteria (open-source software); their criteria for propriety obviously differ widely. Correspondingly, coming closer to the 'truth' is much harder for the latter types of community than for the former. Further, the size of the community involved matters: the larger it becomes, the more tendencies towards stratification and differentiation of power are likely to manifest themselves. Moreover, a multitude of elites may be forming, each with their own agendas.

Keeping these questions in mind let us return to the phenomenon of disruptive contributions and focus on their scale. It can safely be asserted that in the largest of all open-content encyclopedias, Wikipedia, disruption has reached gigantic proportions. For the English language version of the encyclopedia in particular, estimates hover around 9000 malicious edits a day. A few reasons behind this vandalism on a large scale can easily be identified. On the one hand, the Wikipedia corpus involved has assumed large proportions (almost five million entries), making it a very attractive target for visitors of bad intent; on the other, Wikipedia allows co-creation (3.0) with full write-access for all, a collaborative tool that is very susceptible to disruptive actions by mala fide visitors.

From the beginning of Wikipedia, in 2001, several approaches to curb vandalism have been tried. Some of them have been accepted and endured; others have been discarded along the way. In the following, the various tools and approaches are spelled out, with a particular emphasis on the ones that are still being used. It is shown that within Wikipedia a whole collection of anti-vandalism tools has been developed. These instruments are being unfolded in a massive campaign of collective monitoring of edits. This vigilance is exercised continuously, on a 24/7 basis. In the process, human Wikipedians are being mobilized as patrollers, from administrators at the top to ordinary users at the bottom; in addition, several autonomous bots are doing their part. In a short detour, this system of surveillance is analysed from the viewpoint of robotic ethics, and shown to conform to the recent trend of 'coactivity' between humans and bots.

In the central part of this article, subsequently, the moral questions that emerge from this mixed human/bot surveillance are discussed. The analysis is to be seen as an exercise in 'disclosive ethics' (Brey 2000): uncovering features of the anti-vandalism system in place that are largely opaque and appear to be morally neutral—but, in actual fact, are not so. Based on a careful reading of Wikipedian pages that relate to vandalism fighting and a period of participant observation as a practising patroller I draw the following conclusions. As for power, next to administrators, strong anti-vandalism tools are shown to be only distributed to trusted users, thus creating dominance for a new hierarchical layer. As far as surveillance is concerned, filtering of edits becomes questionable when it focuses on editor characteristics like anonymity or reputation. Is Wikipedia engaging in objectionable profiling practices? Further, the system raises the issues of 'over-surveillance' and of the loss of the display of moral skills towards newcomers. Moreover, concerning the institution as a whole, the system of surveillance operates largely as an invisible and opaque technology, hidden from sight to Wikipedian contributors at large. It is argued that as a result, the encyclopedia's governance has become Janus-faced. The wiki platform is an open invitation for all to participate, exemplifying the assumption of trust, but underneath tight surveillance is exercised that starts from the assumption of continuous suspicion. Finally it is argued that in view of these objections the anti-vandalism system needs rebalancing. Also, fair information practices towards contributors need to be developed that highlight the ins and outs of anti-vandalism operations and ask for consent concerning the use of their contributions.

## Anti-vandalism tools in Wikipedia

Basically, vandalism fighting in Wikipedia consists of two phases: edits to entries are *selected* and subsequently *inspected*. Any vandalistic edit that has been detected is corrected by its deletion ('reversion', 'undoing'); usually, an edit comment is attached making mention of vandalism. Performing these steps has always been part and parcel of the write-access permission granted to (almost) all users. The buttons involved that have to be pressed are readily available. This default mode of vandalism fighting—to be denoted the 'basic mode'—soon revealed itself to be insufficient in view of the rising quantity of vandalism. From about 2005 onwards, therefore, a wide range of new tools has been developed by Wikipedians themselves to remedy the situation and achieve more anti-vandalism power. I will not try to give an overview that respects the chronological order in which these were developed; instead, I focus on their basic *functions*. Moreover, I only present a selection of the available tools: those that are most convenient and/or powerful and (therefore) mostly in use.

The first phase of *selection of edits for inspection* has been facilitated as follows. Patrol options have been developed that allow seeing a complete list of recent changes to all entries. The list is refreshed continuously. A common way to obtain these edits is by means of RSS feeds or IRC. In order not to be overwhelmed by this massive flood of edits, *filtering* options have been developed. Three options can be distinguished.

First, one can filter out *types of entries* involved. One may focus on edits in a specific namespace; say the main namespace only (the entries themselves). Further, users may compose a list of specific entries they are interested in (their own personal 'watchlist'), and then select all recent changes to that watchlist only. Another focus that has been enabled is the sensitive category of entries about living people. Secondly, one may choose to focus on the *content* of contributed edits: those with bad words, those with massive blanking, those that delete a whole entry, etc. Blacklists of suspect words and expressions are assembled for this purpose. Thirdly, specific types of *editors* may be targeted. A focus on anonymous contributors, on new accounts, on warned contributors, on users that figure on one's personal 'blacklist'—the possibilities are endless. Similarly, categories of users can be excluded from scrutiny: no administrator edits, no bot edits, no whitelisted users, etc. Further, modern tools allow *combining* various filtering options. For example, a focus on large deletions as committed by IP-accounts can be realized.

After edits have been selected, the second phase of *inspection of edit*s begins. Obviously, any detected vandalistic edit gets reverted. But beyond this action, several *buttons* have been developed that trigger specific actions deemed appropriate. A patroller may easily leave a warning message on the talk page of the supposed vandal, ask for administrator intervention against the supposed vandal ('blocking'), ask for the page to be 'protected' (i.e., lower level users may no longer contribute), propose the entry to be deleted as a whole (even 'speedily'), etc. Another powerful option is a 'rollback' button: after a vandalistic edit has been spotted, it permits not only the reversion of the edit involved but of all antecedent edits to the entry as committed by the same suspected vandal as well, reverting them all in one go. It reverts a whole consecutive series of (supposedly) vandalistic edits, not just the most recent one.

All these options to raise the anti-vandalism powers of Wikipedian patrollers have found their way into various concrete *tools* (for a selection of them, see Table 1). I will mention some of them here. On the #cvn-wp-en freenode channel (accessible via Chatzilla, a Firefox extension), IRC bots (such as SentryBot or CVNBot1) display a continuous stream of edits that are deemed to be suspicious. Moreover, the various reasons for suspicion that apply are mentioned, each reason with its own colour (possible gibberish, large removal, edit by blacklisted editor, edit by IP-account, etc.). Vandal Fighter offers about the same functionalities, though the reasons for suspicion are rendered more succinctly. Lupin is a tool using RSS feeds for displaying recent edits and filtering them by various criteria; subsequently, various buttons are available that come into play after edit reversion. Twinkle installs a menu of buttons on the patroller's screen to facilitate edit correction. The rollback tool, finally, consists of just one button on the patroller's screen for swift reversal (cf. above).

**Table 1** Anti-vandalism tools used in Wikipedia and their affordances beyond the 'basic mode' of fighting vandalism (selection; cf. WP:OLDSCHOOL)

| Phase: | Selection of edits | Inspection of edits |
|---|---|---|
| *Operators with their tools* | | |
| Human operator using Vandal Fighter | Use of filters | |
| Human operator using #cvn-wp-en | Use of filters (also in combination) | |
| Human operator using Lupin | Use of filters | Use of buttons |
| Human operator using Twinkle | | Use of buttons |
| Human operator using rollback | | Use of button |
| Human operator using WPCVN | Use of scoring algorithms | |
| Human operator using Huggle | Use of scoring algorithms | Use of buttons |
| Human operator using STiki | Use of scoring algorithms | Use of buttons |
| Autonomous bot | Use of scoring algorithms | Autonomous action |

*Notes* Before their potential acceptance, edits are filtered by several edit (or abuse) filters; 'basic mode' means that only the basic facilities of the Wikipedian architecture are employed (no additional tools are used); tools mentioned in the table can sometimes usefully be employed together (e.g., Lupin and Twinkle; WPCVN and Twinkle); WPCVN is out of order since January 2014; a recently developed tool, igloo, is not included in the table while still in alpha development

Recently, though, with the tools just mentioned already in place, the fight against vandalism received a giant boost by two more or less independent developments. For one thing, *edit filters* (or *abuse filters)* as a means to prevent intrusion have been developed. These are extensions to the Wikipedian platform that set specific controls on user activities: whenever a user tries to commit an edit, it first has to pass through these filters before ending up in the encyclopedia itself. At the time of writing, after severe testing and discussions, close to a hundred filters have been approved of and are activated on the platform. They can have many a focus, but vandalistic actions are particularly well represented. A particular watch is on the combination of new and/or unregistered users performing such actions. Upon spotting a potentially disruptive edit, various automated actions are possible. The edit may be tagged as potentially vandalistic; such tags can be focussed on by subsequent patrollers. The user may be warned ('are you sure you want to proceed?'), possibly preventing the edit to be submitted at all. The edit may be stopped ('disallow'): it does not pass the filter; the editor is warned in appropriate ways. Finally, filters can have built-in so-called 'throttle': disruptive actions are allowed up to a threshold (say blanking large chunks of text up to once per hour). Passing that threshold triggers the filter into action. Note that most filters in use are public, that is, their functioning can be looked up on a Wikipedian page. Some, however, are kept private and cannot be inspected by the ordinary user. This is done to keep potential vandals in the dark about the possibilities of evading the filter in question (for all remarks in this paragraph, see the links mentioned under WP:Edit filter).

Another recent boost to fighting vandalism, in a carefully tailored way this time, is provided by *computational approaches* to detect vandalism. Research in this area has crystallized into four categories of algorithm (Adler et al. 2011). Algorithms may focus on language features (like bad words, pronoun frequencies), on language-independent textual features (like the use of capitals, changes to numerical content, deletion of text), on metadata of edits (like time and place of the edit, anonymity, absence of revision comment), or on the editor's reputation as a trustworthy contributor. For the latter measure various approaches are in circulation (for reputation as conceived of in the so-called 'Wikitrust model' see Adler and de Alfaro 2007). Computer scientists have been struggling with the question which approach to the detection of vandalism is the most fruitful. Based on a computer tournament with all approaches in the competition the provisional answer seems to be: a combination of all four algorithms works best (Adler et al. 2011).

These vandalism detection algorithms are useful weapons in the struggle against vandalism. They can be enlisted as useful assistants to humans in the phase of selecting edits. If so instructed, algorithms can calculate and assign a probabilistic vandalism score to each and every fresh edit that has passed the abuse filters and comes in. These scores are used to make ordered lists of edits, with the most suspect edits on top. 'Engines' of this kind have found two ways of employment.

On the one hand, they have been incorporated as detection engines in '*assisted editing*' tools. The prime example is the STiki tool (Table 1). At its back-end, all fresh edits from the Wikipedian servers are continuously monitored for vandalism and vandalism scores assigned to them. The method employed in the original implementation was the third, metadata approach; currently, the neural network approach has been enabled also (cf. below). At the front-end, subsequently, using IRC, suspect edits are served to human patrollers in an ordered queue. The filtering task has effectively been taken over by the machine: humans just work through the queue. Edits can be accepted (classified as innocent or pass, the latter option meaning the patroller is not quite sure), or reverted (classified as good-faith revert if no malicious intent seems to be present, or as vandalism if such intent is obvious). Good-faith reversions can be commented on; the diagnosis of vandalism automatically triggers a warning note to be placed on the vandal's talk page. A similar 'assisted editing' tool for the whole process is Huggle (Table 1). Fresh edits are monitored and rated according to a mixture of language, textual and metadata features. Subsequently, in a default queue, these are served to human operators at the front-end, who have a repertoire of actions at their disposal quite similar to the ones described above for STiki.

On the other hand, vandalism detection algorithms have been extended and turned into fully *autonomous bots* (see Table 1) (for a useful overview of the bot landscape in Wikipedia, for several other purposes than anti-vandalism as well, one is referred to Livingstone 2012: 126–132, 180–227 and Nielsen 2014: 17, 35–38). After severe testing they may be let loose on the Wikipedian edit stream. While hundreds of bots are currently in operation, several of them are tailored towards vandalism (another 10–20 of the kind have now been 'retired'). They routinely scan incoming edits for vandalism right after they are published on Wikipedia. A bot intervenes like any human patroller: after identifying a vandalistic edit, it reverts the edit and leaves a warning note on the vandal's talk page. Which kind of algorithm detection is being used? In the beginning most bots relied on the first method: language features. Blacklists (of words) were commonly used. Of late, the four classical methods listed above have (largely) been set aside. Instead, the most recent bots operate as neural networks that gradually learn to distinguish between bad edits and good edits. In order to learn the bots have to be fed continuously with edits as actually classified by humans. Especially this latter type of bot is a promising development. The top scorer among them is ClueBotNG, which can boast of almost three million reversions in total (since 2011). It typically checks edits within 5 seconds of their appearance and reverts about one of them every minute.

## Assessment of the robotic landscape

So the fighting of vandalism on Wikipedia has assumed large proportions, mobilizing literally thousands of volunteers and several bots. Now, how is this amplified process to be understood? Is it business as usual though on a larger scale, or has the process assumed a different character? Are the changes only of a quantitative kind, or also of a qualitative kind?

Geiger and Ribes (2010), elaborating on how the Wikipedian process of 'banning a vandal' has been transformed, firmly take the latter position. They argue that the combined force of humans and bots allows a process of '*distributed cognition*' to unfold, in which collaborators unknown to each other are knitted together in the common purpose of eradicating vandalism. In it, the talk page plays a pivotal role: all warning messages end up there, enabling a coordinated response to ongoing vandalism. So it is only by the creation and deployment of the anti-vandalism tools discussed above that vandalism patrolling becomes possible at all. A sea change has taken place.

I do not quite agree with their diagnosis. Fighting vandalism has been possible in Wikipedia from the beginning. Focussing on specific types of edit or editors has always been possible (though cumbersome); reverting edits while leaving comments about the reversal and/or warning messages on talk pages has always been possible as well (though cumbersome). I want to argue that it is the creation of the Wikipedian platform and the associated wiki tools that must be seen as the cradle for distributed cognition in Wikipedian fashion. The revolution took place in 2001, not around 2011. Of course, the developed tools combined with the large influx of human volunteers have enabled a much larger scale of vandalism fighting. Patrollers may work vastly more efficiently with these tools. Only edits singled out as suspect have to be inspected, only a few buttons have to be pressed for dealing with edits found to be wanting. Without the computational powers involved in particular, vandalism fighting would not have 'scaled' so easily (from watching over a corpus of a handful of entries to almost 5 million entries by now). And without those computational powers, Wikipedia would have been corrupted on a large scale.

Geiger and Ribes (2010:124) also mention '*delegated cognition*': parts of the patrolling task are shifted to computational tools. Edit filtering can be steered by algorithms; edit filtering, inspection, and correction combined can be performed by autonomous bots. It is this delegation that stands out as novel, meriting closer attention indeed. How is this shifting of the burden to be understood?

Let me first remark that the anti-vandalism bots in operation can be interpreted as an instance of 'explicit ethical agents' (Moor 2006). They perform calculations and decide to act depending on the outcome, similar to chess robots. Since they can make decisions on their own, they may exhibit surprising behaviour. If this interpretation is accepted, what about the responsibility of humans vis-à-vis their robotic creations? In what ways are responsibilities for their actions to be distributed between them? In particular, can a *responsibility gap* be detected here (cf. Johnson 2013 for an overview of this issue)? One position is to argue that, since their behaviour is no longer under the control of their creators/operators, the bots involved are to be considered as responsible agents in their own right—a position advanced by Matthias and Sparrow. Another position is to argue that their creators/operators are still to be blamed, either from the point of view of the juridical profession where this is quite normal (Santoro), or on account of their professional responsibility (Nagenborg et al.).[2]

It is interesting to observe that in Wikipedia this issue is answered unequivocally: creators of algorithms and *a fortiori* of bots are held accountable. Lengthy procedures institutionalize this conception. Programmers have to test their new tools extensively in protected trials. Subsequently they have to submit their tools to a committee (Bot Approvals Group, BAG) that asks for logs about the testing and discusses the tools thoroughly on their talk pages. In those discussions the possible damage these tools can inflict on Wikipedian spaces looms large. Letting loose bots that behave erratically or inflict a lot of damage that necessitates elaborate reversal operations is not appreciated. The bot involved will be 'suspended' (for all remarks in this paragraph cf. WP:Bot policy).

From the angle of the *moral design* of robots, an interesting characterization can be made. The vandalism fighting tools that operate autonomously are only allowed to 'act' under very strict and limiting conditions: the threshold of the vandalism probability that triggers action on their part has to be set high with regard to minimizing the false positives rate. The underlying argument is that humans are very sensitive to being falsely reprimanded by a bot. The BAG is keen on minimizing such occurrences. As a result of these limitations, many edits that bots are fairly sure represent vandalism, slip through the net. Hence, much of the anti-vandalism effort continues to rest on human shoulders.

Assisted editing tools are indispensable in that regard. These can be seen to occupy a halfway position between humans with (non-computational) tools and bots. In cases like the employment of Huggle and STiki we observe a fusion between human and computer power. These instruments can readily be interpreted as emanating from the conception of 'coactive design' (Johnson et al. 2010). Such design moves away from making bots more autonomous; instead, its focal point is to make agents more capable of joint activity with interdependent people. In the case of Wikipedia, the sharing reads as follows. Autonomous bots reap the low-hanging vandalistic fruit; subsequently, humans and bots joined in coactivity reach higher for the remaining rotten fruit.

## Relations of work and power

After this overview of the counter-vandalism mechanisms installed, I proceed to investigate questions related to changes in Wikipedian governance as a result. First, how does the revamped monitoring system change the ways in which contributors work together on expanding and refining the Wikipedian repository (relations of work)? To answer this question I rely on a former analysis of mine of the management of trust within open-content communities (de Laat 2014; cf. also de Laat 2010, 2012b).[3] I argued that Wikipedia from the start has opted for a policy of fully open read- and write-access for all. This is an institutional gesture of *trust* towards participants: the gesture signals that they are assumed to be trustworthy in both a moral and an epistemological sense. The introduction of surveillance around the clock for each and every freshly contributed edit significantly qualifies the former gesture. New edits are no longer reviewed casually, as it were, whenever a fellow Wikipedian happens to walk by; now they are put under almost immediate scrutiny. Edits successively have to pass the edit filters installed, survive the swift perusal by autonomous bots like ClueBotNG, and withstand scrutiny from patrollers that preferably use tools like Huggle and STiki. All this is done with one single purpose in mind: keeping Wikipedian namespaces free of damaging contributions. The dominant concern is *damage avoidance*. While before the platform was completely unprotected from vandals, now an army of silicon and human patrollers stands ready to prevent intrusions. Thereby the grant of discretion to ordinary users, as the exercise of one's skills and judgment, has changed. They still have full powers of contributing the contents they wish; full write-access still obtains. At the same time, however, an immediate vandalism check is likely to be performed. As a result, their discretion has been reduced; not by eliminating any of the Wikipedian editing permissions but by much faster performance review.

While keeping this conclusion in mind, another question immediately imposes itself: who actually have obtained the powers to exercise this scrutiny of fresh edits using the anti-vandalism tools available? Who are these gatekeepers who arguably qualify the powers of the ordinary Wikipedian? And, as a consequence, what changes can be observed in the amount of influence on the day-to-day production of entries as exercised by the various levels involved? As it turns out, by no means anyone is entitled to watch the gates with all available tools; on the contrary. The basic rationale underlying the distribution of counter-vandalism tools is the following. Just by pressing some

[2] Precise references to all authors just mentioned can be found in Johnson (2013).

[3] The remainder of this paragraph is a short abstract of de Laat (2014). For full substantiation of the steps in the argumentation that follows one is referred to the article itself.

buttons, these tools may potentially inflict much damage on Wikipedian namespaces. The stronger the tools are, the more harm can be done. Therefore, the stronger tools are basically granted to administrators only, and to those who can prove (as a rule to those same administrators) that they are trustworthy enough for the patrolling task. This line of reasoning applies in particular to tools developed earlier such as the rollback permission, and to tools developed later such as those geared towards assisted editing (Huggle, STiki).

Let me give an indication of the numbers involved. Against a backdrop of millions of ordinary contributors (of which over 22 million have registered), currently about 1400 administrators are active; by default they all have the rollback permission. In addition, they granted that permission to almost 5000 other users (WP:Wikipedians). For the—arguably stronger—Huggle and STiki tools the numbers that have obtained permission to use them are considerably less: currently each tool may be used by about 800 trusted users in total (note that rollbackers obtain the permission by default if they care to ask for it). It turns out that once permission has been obtained, subsequently only some 10–20 % of them actively use the acquired tool(s) over a longer period.

The common Wikipedian, though, has to contend himself with the display of suspect edits through #cvn-wp-en or Lupin, and the filter & button affordances that tools such as Lupin and Twinkle offer; only the *modest* tools in the anti-vandalism Wikipedian repertoire are readily available to those who care to take up patrolling. With this conclusion I challenge, therefore, the observation by Livingstone (2012: 213) that "as watching functions are […] available to all users, the control element of surveillance remains largely distributed across the [Wikipedian] site." He simply overlooks the harsh requirements (in terms of editing experience) for obtaining effective patrolling tools.

In sum, institutional trust to watch on fresh edits (as a specific kind of editing) extends to anyone—but the sophisticated heavy instrumentation that effective watch requires is only entrusted to the selected few that have actually proved to be trustworthy members of the community. Can this arrangement be justified? Of course, administrators could no longer handle the vandalism problem on their own. They were in need of more eyeballs to watch the bugs (i.e., vandalistic edits) injected into the Wikipedian corpus of entries (and beyond). They then chose to recruit a special police force, and arm them with dedicated and powerful weapons. This, of course, created a new layer in the otherwise rather flat Wikipedian hierarchy, basically only consisting of administrators and common users. Was the disturbance of the valuable asset of largely egalitarian relations a necessity in view of the struggle against vandalism? Can the unbalancing of power relations be justified by the argument that the integrity of Wikipedia is to be maintained? At the end I return to these questions, which allows us to take other factors into consideration as well, such as several troubling aspects of the implemented surveillance and the opaqueness of the anti-vandalism system as a whole (to be discussed).

## Surveillance

After this overview of the tools against vandalism and their users, it is time to ask the question: does the surveillance in place respect moral intuitions? In the following I maintain that both the use of editing tools by humans and the practices of bots raise serious ethical concerns.

### Profiling

A first problem emanates from the phase of selecting possible candidates for vandalism. Sorting out bad edits may proceed on the basis of *edit* characteristics: language features, textual features, and the like. Whether incorporated in filters or full blown algorithms, the practice seems unproblematic: after all, vandalism detection is all about spotting non-appropriate textual edits. Sorting, however, can also proceed on the basis of *editor* characteristics. Whenever these relate to the user's past behaviour in regard to vandalism (being warned before, on a blacklist, and the like), or to the user's present behaviour arousing suspicion (e.g., omitting an edit commentary when this reasonably seems indicated), the method is uncontroversial. I would argue that even an enabled focus on new accounts can be defended: full credit as a trustworthy contributor does not have to be given straight away. But a few other editor characteristics used for closer inspection *do* seem controversial.

What are we to think of a criterion such as being a *non-registered* user (anonymous, with IP-account only)? Filtering out anonymous contributors is enabled in many a tool (Vandal Fighter, #cvn-wp-en, and Lupin). Confronted with an avalanche of fresh edits on the screen a patroller may easily make a choice and focus specifically on anonymous users (such edits are usually indicated by a special colour). For the dedicated chaser of 'anons' (as anonymous contributors are called in Wikipedian parlance) there is even a website that exclusively displays anonymous edits in realtime from all over the world, with the precise country where they are committed (RCMap at http://rcmap.hatnote.com; not in Table 1); with one click they pop up on one's screen ready for inspection. In all these instances the choosing of edits submitted by IP-accounts is a very alluring option to a patroller in view of the returns it may bring (since such accounts are known to be more vandalism-prone).

In a more sophisticated fashion, several assisted editing tools (Huggle, STiki) offer their operators a queue of edits that have already been filtered and ordered according to an algorithm that, besides many other factors, includes anonymity as a warning sign. So routinely, for the designers of these tools, being an unregistered user triggers extra suspicion and generates an increase in surveillance. There can even be a subtle *multiplier* effect at work here, specifically for these tools. When using the priority queue in Huggle or STiki (or WPCVN as well), operators can actually *see* whether the suspect edit on top of the queue is performed anonymously or not, and may—whether unconsciously or not—proceed to give it extra scrutiny. In these instances, effectively, anonymous edits are given a *more severe* check.

Let me stress that it is not farfetched to assume that human patrollers, whether using simpler tools like Vandal Fighter or more sophisticated assisted editing tools, will indeed choose to weigh the odds against anonymous users. In many discussions about vandalism, considerable aggression is ventilated against supposed vandals, and anonymous users are often depicted as being almost synonymous with vandals (cf. various quotes in de Laat 2012a).

Now, what is the problem I want to stipulate here? All practices of filtering as just described are instances of *profiling*: an ensemble of dimensions is bundled together into what is called a profile that by design based on past experience yields higher chances of catching vandals than just random screening (for an elaborate discussion of profiling cf. Schauer 2003). Such profiling is understandable and justifiable, given the quest for optimal detection of vandalism using scarce resources. In many of these profiles, however, whether consisting of just one or of many dimensions, anonymity figures as an important feature. This is simply due to the fact that anonymous contributors (contributing about 15 % of all edits) demonstrably commit much more vandalism than users operating from a registered account; including anonymity in a profile then produces more 'hits' and eliminates more instances of vandalism. Nevertheless, such profiling on anonymity seems questionable, since it risks enhancing stigmatisation of contributors who for one reason or another choose to remain anonymous. Anonymous users turn into a category to be treated with ever growing suspicion.[4]

Many of us condemn our police forces for relying on profiles that include the criterion of race and our custom authorities at airports for using profiles that include ethnic and religious criteria. Incorporating such 'sensitive' dimensions, the argument goes, can only aggravate existing social tensions. Shouldn't we similarly condemn profiling along the lines of anonymity in the Wikipedian case?

A comparison with the Turnitin plagiarism detection system as discussed by Vanacker (2011) may be useful here. Often, students in class are required to submit their papers to the system, in order to preclude plagiarism. Is such a practice permissible the author asks? He answers that it is, since the checks are to the benefit of the whole educational system and uphold the value of their certificates. This would be otherwise, he remarks, if only "student athletes or transfer students" were to be singled out for inspection of their papers (idem: 329). I subscribe to his position, in view of the divisions that would otherwise be created. And on condition that we accept the analogy between plagiarism and vandalism, we are obliged to extend the same stance to Wikipedia: profiling along the lines of anonymity, although purportedly contributing to more efficient detection of vandalism, is to be avoided since it relays the message that anonymous contributors are up to no good.

Some more questionable profiling is taking place, though on a minor scale compared to the focused attention on edits from IP-accounts. The time and place an edit was made are also singled out as criteria for increased vigilance: the metadata-based queue in STiki uses a classifier that incorporates these data. The rationale is that vandalism appears to occur more regularly during specific time intervals (between 8AM and 8PM, and during weekdays as opposed to weekends), yielding time-of-day and day-of-the-week as indicators. Further, vandalism to English Wikipedia has been observed to be more prevalent among American (as well as Canadian and Australian) users than among European (Italian, French, or German) users; so being American turns into a warning sign (for both kinds of data cf. West et al. 2010). In combination, therefore, we observe the following, STiki-specific profiling: being an American contributing to Wikipedia during regular work hours is raising suspicion per se and triggers increased vigilance.

## Reputation

Further, the use of a reputational measure is questionable. For the moment, it has only been used as a scoring tool in assisted editing (STiki), not in bots since its incorporation in them appears problematic. The measure orders the queue that is displayed to its human operators. Apart from being a very complex and very expensive undertaking since by definition it amounts to real time computing, the concept of reputation itself seems hard to operationalize. Based on the Wikitrust model a measure for reputation has been proposed, roughly the sum total of edits by a specific contributor that have survived the testing by subsequent editors

[4] Note that anonymous accounts are not necessarily new accounts (a factor just mentioned as a justifiable element in a profile). Many IP-accounts belong to institutions (educational or otherwise) that grant their members collective access to Wikipedia. Vandalism may then erupt at the first edit, but just as well only many edits later, whenever a malevolent individual member of the institution decides to access Wikipedia and commit a vandalistic edit.

and remained intact. Since then, a whole literature has been developing about the issue, and ever more subtle measures are being proposed.

Two tricky issues, though, are connected with the reputational measure (for both issues cf. de Laat 2014; based on West et al. 2012, Adler and de Alfaro 2007). In order to explain this, its meaning and functions more generally first have to be spelled out. Reputation is not a particular characteristic of an editor at the moment that (s)he submits a specific edit (as in other vandalism algorithms), but it indicates a current *summary* of all of someone's achievements so far in the community. It is intended to be a measure of what can be expected from him or her. As such it can be used more broadly than just as an indicator for possible vandalism: to motivate members to continue performing, to regulate the distribution of editing privileges, to promote contributors to higher ranks, etc. Used in these ways, reputation becomes decisive for one's *whole career* in the community (cf. de Laat 2014).

After this short detour the first tricky issue connected with using reputation as an intelligent routing tool for detecting vandalism can be spelled out: is the measure to be made publicly available or is reputation to be tracked in silence? The first option has much to say for it. For one thing, it would satisfy principles of openness and transparency, which are clearly desirable given the broad importance of the measure for one's life chances in the institution. For another, it allows reputation to function as an incentive for proper behaviour – much like in eBay. As mentioned above, seeing one's efforts reflected in (higher) reputation is supposed to be stimulating. Unfortunately, such visibility would at the same time invite 'gaming' the system in various ways (such as by dividing a contribution into smaller consecutive edits) and thereby undermine the measure's accuracy. As yet, no solutions to such vulnerabilities have been found. Switching to the alternative of tracking reputation in silence, then, is the other option—similar to how things are presently done in some corners of Wikipedia. However, although gaming the system is no longer an issue, any incentive for proper behaviour would be eliminated. More importantly, transparency would be forfeited. So one way or another, efficiency and morality always seem to clash here: either the measure conforms to moral standards but is not very efficient (public reputation), or the measure is reasonably accurate but violates moral intuitions considerably (secret reputation).[5]

The second thorny issue connected with any measure of reputation is a technical issue: which starting value is to be chosen? Putting it in short terms: given its range (say from 0 to 1) a starting value at the middle (1/2) is preferable, with this value going up and down as a contributor's actions unfold. Unfortunately, this allows vandals after 'bankruptcy' to start all over again from a new account. So this option does not deter any vandalism. Starting at value zero avoids this problem; but then newcomers and vandals receive the same kind of vigilance, which is clearly suboptimal (as well as slightly immoral).

In view of these problems, any measure of reputation is bound to be problematic. Its use as the basis for queuing would amount to a capricious and haphazard surveillance practice. No wonder, that after some try outs on dumps from Wikipedia for research purposes, apart from use in WPCVN and STiki, as of today reputational algorithms are nowhere in use in Wikipedia.

## Oversurveillance

Yet another moral issue has to do with a tendency towards overuse. As anti-vandalism tools become stronger and more sophisticated, they are more easily prone to be overused; as a result, quantity trumps quality, defeating the original purpose of these tools. Let me explain.

First consider assisted editing tools like Huggle and STiki. Since they display new edits in an instant (the ever present priority queue), they represent an invitation to treat edits ever faster. One may always press the buttons involved faster and faster. The practice turns into a computer game that may properly be called 'shooting the vandals'. Unwittingly or not, STiki in particular is specifically stimulating such practice, by maintaining daily leader boards of STiki editors that display the amount of edits treated by them and the actual reversion scores obtained (WP:STIKI/L). 'Gamification' has taken hold of the patrolling domain. Some troubling figures emerge from these leader boards. Some patrollers treat hundreds of suspect edits a day (on average). Actual reversion scores on account of vandalism vary wildly between STiki patrollers: from a modest 5 % up to 80 and 90 %. And it has to be borne in mind in this regard that the edit queue cannot be chosen here; there is always just one obligatory next edit a patroller cannot escape from. I would argue that the figures yield reasonable indications of overuse of the tool by at least some patrollers.[6]

A parallel tendency applies to bots. Bots too can be turned into overzealous patrollers. That is to say, their parameters for action can be tuned in order to increase the catch of vandalistic edits; whether innocent edits are

[5] Note that I come back to the transparency vs. obscurity issue below, since it pertains to the anti-vandalism system as a whole, not just to the use of reputation in an engine for detecting vandalism.

[6] Playing games is the province of men rather than women. Would the tendency toward gamification signalled above by any chance reflect or even reinforce the current male predominance in the Wikipedian population as a whole? (Thanks for this suggestion are due to an anonymous reviewer of this journal).

eliminated in the process matters less. Or in technical terms: the tracking rate ('recall') is set ever higher, at the expense of the false positives rate also rising. It has to be mentioned here that this problem of fine-tuning has not gone unnoticed in Wikipedia. The Bot Approvals Group has discussed the issue at length, and maintains a rule that bots should have a false positives rate at most equal to the rate that humans achieve in practice.[7] A bot should perform better than humans, in the sense of hardly bothering innocent editors—not so much in terms of the number of vandals caught.

With both humans and bots potentially overstretching their reach, one should not lose sight of the fact that the basic algorithms underlying their patrolling actions are often generated by machine learning. The rationale of such results is inherently opaque: neural networks just learn in their machine way. The results can be surprising, even to their creators. I have no specific evidence of this happening, but it is an aspect worth keeping in mind.

### Bots taking over

Over the last few years, there has been a notable tendency of bots taking over the chores of vandalism fighting from humans. According to recent estimates bots already eliminate half of all vandalistic contributions. Numerous are the messages from human patrollers on the bots' talk pages, congratulating them on their speed and accuracy. This advance of the bots has several troubling aspects.

First, patrolling against vandalism is not only a technical task, of adequately identifying a vandalistic edit and taking appropriate action. If done well, it also demands the exercise of *moral skills* that make up the character of the virtuous patroller (for the introduction of this term in the context of new technology cf. Vallor 2015). These have to do with displaying restraint, with being tactful, forgiving, and supportive in gentle ways. Not all supposed vandals are what they seem to be. By exercising these skills, well-meaning newcomers may be saved for the Wikipedian project, instead of being pushed away, never to come back. With bots taking over, then, we may ask whether the required moral skills are exercised at all. After all, the present bots just leave behind preformatted templates, delivered and signed by a name ending with BOT. Such a treatment can hardly be considered inviting. It has been observed, similarly, that already one-third of newcomers to Wikipedia obtain their first return messages from a bot. Significantly, many human users 'caught' by an anti-vandalism bot are reported as trying to talk back to the bot—overlooking the bot label.

Secondly, in line with the foregoing, if bots really take over from humans in the future, the associated moral skills may well be lost among human patrollers themselves: such skills erode when not in use. Patrolling vandals is—and should be—a continuous practice in restraint and diplomacy. That kind of schooling is in danger of fading away. All that remains then is the metal tone of the bots scouting Wikipedia on their own.

Bots taking over, though, is not really the policy at Wikipedia. Out of fear of alienating too many potential contributors, bots may only revert the very high probabilities; any edits scoring lower are left for human patrollers. 'Coactivity' of algorithmic power and humans is the current policy (cf. above).

This is a fortunate trend. After all, what *is* vandalism, let alone *obvious* vandalism? An edit *per se* can (almost) never be labelled vandalism. Any change of numbers; any change of names; any change of web link; any paragraphs added or deleted—it all depends on the context as to whether this is vandalism or not. And associated with this, any vandalism can only be obvious to a patroller who is familiar with the context. Let me quote from my own experience. Changing the number of Jews murdered in WWII from 6 million to 6 thousand is an obvious bad faith vandalistic change to me—but not every patroller is familiar with WWII. The number change I came across that took place in the sales figures of a particular company turned out to be vandalism as well. But this was not obvious to me since I am not acquainted with that type of company; I had to do some research before reaching my verdict. Similarly, the deletion of several paragraphs in a specific entry without leaving a comment seemed obvious vandalism to me at first; any other patroller would draw the same conclusion. Then it emerged that the editor involved had been talking it out on the associated talk page; the deletions were part of a consensus for action reached there. (S)he protested and I had to apologize and revert my own reversion of the deletion. These musings serve to underline that the trend in Wikipedia to include both bots and humans in patrolling is a fortunate one indeed—vandalism detection can never be automatic and foolproof at the same time.

### Wikipedia as a Janus-faced institution

The—to my view—most alarming aspect of the Wikipedian anti-vandalism system as a whole is the air of secrecy and opaqueness in which it operates. In order to develop this argument it is most appropriate to portray the encyclopedia as a *Janus-faced institution.* As Johnson

[7] Human Wikipedians roughly achieve a false positives rate of 0.25 %: one in 400 legitimate edits is mistakenly classified as a vandalistic edit. ClueBotNG initially started off with the same false positives rate. By now, this rate has been lowered to 0.1 %; that is, at most one in 1000 legitimate edits is mistakenly identified as a vandalistic edit (WP:CLUEBOT).

(2004) has argued, evolving ICT impinges on the instrumentation of human action, by creating new ways for doing the things we can do already as well as possibilities for doing novel things. In the case of Wikipedia the advance of the associated technologies has resulted in a particular instrumentation of human action: the tools for human action may be interpreted as having *bifurcated* into two contrasting types. As a result, the institution has acquired two quite opposed faces. The contrast may usefully be drawn out by taking recourse to the distinction between *transparent* technologies on the one hand and *opaque* technologies on the other, as elaborated by Lucas Introna in the context of the politics of surveillance cameras (Introna 2005).

The basic platform of Wikipedia (an implementation of MediaWiki software), consisting of main pages, talk pages and the like, presents itself in an accessible way to all users alike. They are invited to become involved and assist in developing entries. Further, all contributions are logged and publicly available. So *transparency* can be said to reign. Things are quite otherwise for the whole array of anti-vandalism tools, whether humanly operated or fully autonomous bots. For one thing, these tools are normally hidden from view: as long as co-creative editing goes smoothly, nobody will notice that checking for vandalism is taking place continuously. Only when humans or bots start to interfere by carrying out reverts and/or leaving warning templates, will participants become aware of the patrolling in the background. As well, the patrolling is a form of surveillance that does not need any extra involvement from the surveilled; they can just carry on their editing. Finally, and most importantly, operation and outcomes of the tools are most obscure. The operation is steered by a variety of algorithms; the more sophisticated they become, the more opaque they are (cf. metadata algorithms and neural networks). They may properly be considered black boxes in their own right. So I argue that taken together the set of anti-vandalism tools in use can aptly be denominated an *opaque* technology. The Wikipedian technologies involved arguably follow closely the distinctions laid out by Introna. It is worth noticing here that earlier I denominated the second face a 'background' mechanism (de Laat 2012b), and Geiger (2011) referred to it as the 'hidden order' of Wikipedia—though he mainly had bots in view.

Another difference between the two Janus faces has to do with the *purposes* of the technology involved. The Wikipedian platform on the surface exemplifies the invitation for all to collaborate on entries and let the encyclopedia prosper. "We trust you all" is the signal it emits. The array of surveillance practices below the surface, however, is the embodiment of suspicion. Edits from (almost) all contributors are systematically checked, behind their backs, in order to avoid possible damage. The system of surveillance signals: "You are to be watched closely." This contrast, added to the transparent-opaque distinction drawn by Introna, only serves to accentuate the tension between the two faces.

Remarkably, the two faces are also decoupled *technologically*. The MediaWiki platform on the one hand and the anti-vandalism tools on the other are not a technologically integrated system. The latter tools are simply a collection of browser extensions, standalone programs, and bots, installed on and operated by the individual patrollers' computers. While not running on the Wikipedian servers, they do not impede the platform's performance. Geiger dubs this 'bespoke code' and estimates that the number of lines of code involved in all of them together is higher than the size of the code base of the MediaWiki platform itself (Geiger 2014). Notice though that edit filters (cf. above) as part of the anti-vandalism system are the exception to this rule: they run directly on the Wikipedian platform.

This contrast between the transparent wiki face and the opaque anti-vandalism patrolling face of Wikipedia is related to a dichotomy signalled by Stegbauer (2011). His thesis is that the encyclopedia's ideology is gradually changing from an 'emancipation ideology' which stresses that everybody is welcome to contribute, to a 'production ideology' emphasizing that high-grade entries have to be produced. The former ideology developed in Wikipedia's earlier years; later on, with the size of the encyclopedia growing beyond comprehension, the latter ideology took hold. It became embodied in the appointment of administrators with the powers to protect pages and block users: they keep a watchful eye over vandals, trolls, IPs, and difficult and unreasonable people in general. As a result, as of now, newcomers have an ambiguous experience. The emancipation ideology entices them to participate ('the encyclopedia that everybody can edit'); subsequently, they have the sobering experience of a myriad of rules they have to follow and a range of functionaries they have to obey. Notice in this regard that in my interpretation of the recent discussion among Wikipedians about a system of reviewing edits before they appear on the screen (the so-called flagged-revisions scheme) a similar dichotomy between 'process' and 'product' ideology surfaced (cf. de Laat 2012a).[8] The connection of the Stegbauer thesis with the growth of anti-vandalism tools and the recruitment of patrollers in particular may be clear: such patrolling is only the latest step in the expansion of the hold of the 'product ideology'.

This Stegbauer thesis about Wikipedia is an exciting one. It is tempting to go one step further and interpret it as a particular instance of the Iron Law of Oligarchy as

[8] For a broad overview of the bureaucratic problems facing Wikipedia see Simonite (2013). In the article these problems are argued to be (partly) responsible for the declining number of editors and the continuing lack of editor diversity (predominantly male, technologically-minded, and from the Western hemisphere).

formulated by Robert Michels in 1911. According to that law, all organisations, whether or not of democratic origin—like trade unions and political parties, tend to evolve towards oligarchy. Would that law by any chance also hold for web-based open content communities—Wikipedia in particular?

Obviously, this opaqueness of the anti-vandalistic face of the Wikipedian institution goes against the basic principles of transparency and accountability. Wikipedia, in particular, as a site that projects itself as an encyclopedia for all, cannot allow itself to just slide away, step by step, into developing surveillance practices in the background that operate ever so silently and opaquely. This threatens to erode Wikipedia's moral order as a whole. In a democratic institution such practices cry out for clarification and justification, for the exercise of democratic control.

## Conclusions

It has been argued that in Wikipedia surveillance in order to fight vandalism has assumed morally questionable proportions. In the process, some users have established themselves as more powerful actors than others. Profiling of anonymous users has become a routine procedure; questionable measures of reputation have also come under consideration for surveillance purposes. Further, overzealous patrollers and/or bots may become a nuisance for good faith contributors. As well, the ever needed moral skills in 'treating' vandalism are in danger of being eroded. Finally, an air of secrecy and opaqueness surrounds the whole patrolling venture, in particular as a result of the anti-vandalism algorithms and neural networks in use that are opaque to all concerned—ordinary users and experts alike. All of this is in need of public discussion.

Some may argue at this point (or already earlier on), that mystery or no mystery, vandalism is an evil that may destroy the reliability of Wikipedia as an encyclopedia for all. Reliability that has been built up gradually is in danger of being eroded. Of course I can only agree with this diagnosis—without considerable anti-vandalism efforts the encyclopedia would have been doomed to failure a long time ago already. The whole point is, however, whether this threat alone is enough to force us to swallow any amount and any type of such activities unfolding, thereby effectively closing the discussion. I would argue that instead a more nuanced and extended discussion should take place about the pros and cons of the several aspects of surveillance. We should seek to acquire a better balance concerning the various issues involved.

Such discussion, however, is severely hampered precisely by the opaqueness in which the whole anti-vandalism technological apparatus is clouded. That mystery is not only objectionable in itself; it likewise hampers public discussion developing about the pros and cons of such patrolling. Similar observations about 'obscure' technologies as immune to public scrutiny can be found in Introna (2005) concerning surveillance cameras, and in Stahl et al. (2010) concerning computer security and computer forensics. It is also the same kind of conclusion as Friedland (2014), a lawyer, writing *after* the Snowden revelations, draws concerning modern day surveillance practices and privacy. He argues that these practices become more and more invisible. Thus, the targets of surveillance (i.e., all of us) have no knowledge of being surveilled and see no reason to raise their voices. As a result, the necessary exercise of their democratic rights of speech is thwarted. Only transparency can do justice to the intentions as worded in the US constitution, and enable a fine-tuning of checks and balances concerning privacy in this age of omni-present surveillance.

The foregoing comparison is not in any way intended to suggest that the hidden face of Wikipedia is of comparable importance or causes similar harm as the hidden face(s) of current surveillance practices—such is obviously not the case. Nor is this to suggest that Wikipedia as an institution operates as mysteriously and in the dark as the secret services of the Western nations. At various spots on Wikipedia details of its practices of collective monitoring are documented and even discussed—absolute secrecy or confidentiality does not apply. If they persist, diligent Wikipedian users can take up their accountability in this and piece together information about the anti-vandalism tools that are operative—at least if they have the time and energy to invest in the undertaking. A serious obstacle to overcome though is the circumstance that such information is very much scattered all over Wikipedian namespaces.[9]

In order for this discussion to unfold I would argue that Wikipedian functionaries (whether paid or unpaid) have a special responsibility—after all, they are running the day-to-day affairs of the encyclopedic site. In taking up this responsibility they must as it were put effort into reverting any (natural) trend towards oligarchy—thereby testifying to the fact that Robert Michels' law may not be an iron law after all. One cannot expect from an ordinary Wikipedian that he/she takes the lead in such matters. I have to grant that many discussions about various such issues are already being conducted all over the place—but it is mainly the seasoned Wikipedians who know where to find them and how to take part in them. The enlargement of the base of discussants and

[9] The hardest nut to crack is the opaqueness of the *algorithms* in use: the inner workings of tools like Huggle and STiki, as well as autonomous bots, can only be grasped fully by actually putting the tools to use and seeing what happens. For that reason I have invested time in actually patrolling Wikipedia with them and finding out the intricate details involved.

the mutual coordination of—now scattered—discussions about (algorithmic) transparency would seem to be imperative.

At any rate, whatever the outcome of this rebalancing discussion, the Wikipedian community has to develop more ethical information practices towards its contributors. As of now they are cordially invited to contribute and enjoy the collaborative process—be bold is the motto. In the process, they are urged to remain polite, to do no harm, and to respect copyright laws. Consent (implicit, informed consent that is) is only sought for handing over their edits with a Creative Commons License 3.0 (CC-BY-SA), which stipulates that other users may read, distribute, and modify the edit (while attaching the same license again) (for all the above see Wmf:Terms of Use). But no consent is sought for the ways in which their personal edits are used and processed subsequently; only some vague allusions to this can be found in the privacy policy that applies (Wmf:Privacy Policy).

A comparison might be useful here. In his discussion about the morality of using Turnitin software against plagiarism, Vanacker (2011) argues that 'fair information principles' would require the institution to develop a 'code of ethics' for instructors in the classroom. His discussion, however, mainly revolves around the use of personal data and concerns of privacy. Our case, however, is neither about personal data (in the strict sense), nor about privacy. It is about the several ways in which the data about their personal edits are employed for anti-vandalism purposes. Accordingly, contributors have to be warned about this. They have to be made aware that their edits are routinely surveilled in order to detect vandalism, by a multitude of human patrollers and autonomous bots. Mistakes, they must be told, are inevitable (false positives). Moreover, they should be alerted to the fact that secondary use is involved as well. Both edit and editor data are used as input for several machine learning tools associated with Wikipedia, can be aggregated into a measure of reputation, and—a point not mentioned before—can be downloaded by any computer scientist who wants to analyse dumps of Wikipedian data for research purposes. A whole spectrum of secondary use of information is at stake.

Potential editors, then, must be presented an explicit choice: a basic opt-in that grants consent to the anti-vandalism practices employed, a more expanded opt-in that grants consent to secondary uses as well—or refrain from participation. Other options (such as just submitting one's edit-as-is, without allowing any further processing) are simply not feasible.

## References

All websites below have last been accessed on December 29, 2014.

Adler, B. T., & de Alfaro, L. (2007). A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th international conference on World Wide Web*, May 8–12, 2007, Banff, Alberta, Canada. doi:10.1145/1242572.1242608.

Adler, B. T., de Alfaro, L., Mola-Velasco, S. M., Rosso, P., & West, A. G. (2011). Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *CICLing '11: Proceedings of the 12th international conference on intelligent text processing and computational linguistics, LNCS 6609* (pp. 277–288). Tokyo, Japan.

Brey, P. (2000). Disclosive computer ethics. *Computers and Society, 30*(4), 10–16.

Chen, A. (2014). The laborers who keep dick pics and beheadings out of your facebook feed. *Wired* October 23. http://www.wired.com/2014/10/content-moderation/.

de Laat, P. B. (2010). How can contributors to open-source communities be trusted? On the assumption, inference, and substitution of trust. *Ethics and Information Technology, 12*(4), 327–341.

de Laat, P. B. (2012a). Coercion or empowerment? Moderation of content in Wikipedia as 'essentially contested' bureaucratic rules. *Ethics and Information Technology, 14*(2), 123–135.

de Laat, P. B. (2012b). Navigating between chaos and bureaucracy: Backgrounding trust in open content communities. In K. Aberer, et al. (Eds.), *Social informatics. 4th International conference, SocInfo 2012; Lausanne, Switzerland, 5–7 December, 2012; Proceedings*, *LNCS 7710* (pp. 534–557). Heidelberg: Springer.

de Laat, P. B. (2014). From open-source software to Wikipedia: 'Backgrounding' trust by collective monitoring and reputation tracking. *Ethics and Information Technology, 16*(2), 157–169.

Dutton, W. H. (2008). The wisdom of collaborative network organizations: Capturing the value of networked individuals. *Prometheus, 26*(3), 211–230.

Forte, A., & Lampe, C. (2013). Defining, understanding, and supporting open collaboration: Lessons From the literature. *American Behavioral Scientist, 57*(5), 535–547.

Friedland, S. (2014). The difference between invisible and visible surveillance in a mass surveillance world. Elon University Law Legal Studies Research Paper No. 2014-02. Available at SSRN. http://ssrn.com/abstract=2392489.

Geiger, R. S. (2011). (2011) The Lives of Bots. In G. Lovink & N. Tkacz (Eds.), *Critical point of view: A Wikipedia reader* (pp. 78–93). Amsterdam: Institute of Network Cultures.

Geiger, R. S. (2014). Bots, bespoke, code and the materiality of software platforms. *Information, Communication & Society, 17*(3), 342–356.

Geiger, R. S., & Ribes, D. (2010). The work of sustaining order in Wikipedia: The banning of a vandal. In *CSCW 2010*, February 6–10, 2010, Savannah, Georgia.

Introna, L. D. (2005). Disclosive ethics and information technology: Disclosing facial recognition systems. *Ethics and Information Technology, 7*(2), 75–86.

Johnson, D. G. (2004). Computer ethics. In L. Floridi (Ed.), *The Blackwell guide to the philosophy of computing and information* (pp. 65–75). Hoboken, NJ: Wiley.

Johnson, D. G. (2013). Negotiating responsibility in the discourse on moral robots and other artificial agents. In: E. Buchanan, P. B. de Laat & H. T. Tavani (Eds.), *Ambiguous technologies: Philosophical issues, practical solutions, human nature. Proceedings* *of the tenth international conference on computer ethics—Philosophical Enquiry*, 2013 (pp. 182–193). Lisbon, Portugal.

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C., Sierhuis, M., & van Riemsdijk, B. (2010). Toward coactivity. In: *HRI '10: Proceedings of the 5th ACM/IEEE international conference on human–robot interaction* (March 2010).

Livingstone, R. M. (2012). *Network of knowledge: Wikipedia as a sociotechnical system of intelligence*. Dissertation, University of Oregon, September 2012.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*(4), 18–21.

Nielsen, F. A. (2014) Wikipedia research and tools: Review and comments. Working paper. Most recent update obtained from http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6012. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2129874.

Schauer, F. (2003). *Profiles, probabilities, and stereotypes*. Cambridge, MA: The Belknap Press of Harvard University Press.

Simonite, T. (2013). The decline of Wikipedia. *MIT Technology Review, 116*(6), 51–56.

Stahl, B., Elizondo, D., Carroll-Mayer, M., Zheng, Y., & Wakunuma, K. (2010). Ethical and legal issues of the use of computational intelligence techniques in computer security and computer forensics. In: *Proceedings of the 2010 international joint conference on neural networks*, July 18–23, 2010, Barcelona, Spain.

Stegbauer, C. (an interview with). (2011). Cultural transformations in Wikipedia or 'from Emancipation to Product Ideology'. In G. Lovink & N. Tkacz (Eds.), *Critical point of view: A Wikipedia reader* (pp. 342–350). Amsterdam: Institute of Network Cultures.

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology, 28*, 107–124.

Vanacker, B. (2011). Returning students' right to access, choice and notice: A proposed code of ethics for instructors using Turnitin. *Ethics and Information Technology, 13*(4), 327–338.

West, A. G., Chang, J., Venkatasubramanian, K. K., & Lee, I. (2012). Trust in collaborative web applications. *Future Generation Computer Systems, 28*(8), 1238–1251. doi:10.1016/j.future.2011.02.007.

West, A. G., Kannan, S., & Lee, I. (2010) Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata. In *EUROSEC'10*: 22–28, Paris, France.

## Wikipedia-related references

Irc://freenode/cvn-wp-en, opens in Chatzilla (an IRC-client to be enabled on Firefox).

Wmf:Privacy Policy. https://wikimediafoundation.org/wiki/Privacy_policy

Wmf:Terms of Use. https://wikimediafoundation.org/wiki/Terms_of_Use

WP:Bot policy. https://en.wikipedia.org/wiki/Wikipedia:Bot_policy

WP:CLUEBOT. https://en.wikipedia.org/wiki/User:ClueBot_NG

WP:Edit filter. Combined information obtained from https://en.wikipedia.org/wiki/Wikipedia:Edit_filter, https://en.wikipedia.org/wiki/Special:AbuseFilter, and https://en.wikipedia.org/wiki/Special:Tags

WP:Huggle. http://en.wikipedia.org/wiki/Wikipedia:Huggle

WP:Lupin. https://en.wikipedia.org/wiki/User:Lupin/Anti-vandal_tool

WP:OLDSCHOOL. http://en.wikipedia.org/wiki/Wikipedia:Cleaning_up_vandalism/Tools

WP:STiki. https://en.wikipedia.org/wiki/Wikipedia:STiki

WP:STIKI/L. https://en.wikipedia.org/wiki/Wikipedia:STiki/leaderboard

WP:Twinkle. https://en.wikipedia.org/wiki/Wikipedia:Twinkle

WP:VandalFighter. https://en.wikipedia.org/wiki/Wikipedia:VandalFighter

WP:Wikipedians. https://en.wikipedia.org/wiki/Wikipedia:Wikipedians