Ethics Inf Technol (2016) 18:131–148
DOI 10.1007/s10676-016-9399-8

ORIGINAL PAPER

Profiling vandalism in Wikipedia: A Schauerian approach to justification

Paul B. de Laat[1]

Published online: 30 April 2016

Abstract In order to fight massive vandalism the English-language Wikipedia has developed a system of surveillance which is carried out by humans and bots, supported by various tools. Central to the selection of edits for inspection is the process of using filters or profiles. Can this profiling be justified? On the basis of a careful reading of Frederick Schauer's books about rules in general (1991) and profiling in particular (2003) I arrive at several conclusions. The effectiveness, efficiency, and risk-aversion of edit selection all greatly increase as a result. The argument for increasing predictability suggests making all details of profiling manifestly public. Also, a wider distribution of the more sophisticated anti-vandalism tools seems indicated. As to the specific dimensions used in profiling, several critical remarks are developed. When patrollers use 'assisted editing' tools, severe 'overuse' of several features (anonymity, warned before) is a definite possibility, undermining profile efficacy. The easy remedy suggested is to render all of them invisible on the interfaces as displayed to patrollers. Finally, concerning not only assisted editing tools but tools against vandalism generally, it is argued that the anonymity feature is a sensitive category: anons have been in dispute for a long time (while being more prone to vandalism). Targeting them as a special category violates the social contract upon which Wikipedia is based. The feature is therefore a candidate for mandatory 'underuse': it should be banned from all anti-vandalism filters and profiling algorithms, and no longer be visible as a special edit trait.

Keywords Algorithms · Bots · Discrimination · Profiling · Rules · Vandalism · Wikipedia

> *[I]t is the very silence itself, the ability to take things off the agenda as well as to put them on, that explains much what is valuable about rules.*
> (Schauer 1991: closing sentence on p. 233)

Introduction

So-called open content communities thrive on the contributions from their respective crowds in order to produce software, news, reference entries, videos, maps, and the like. Well-known examples include Linux, Reddit, NowPublic, Wikipedia, and YouTube. Basic parameters for communities of the kind are twofold (cf. Dutton 2008; de Laat 2012). On the one hand we must distinguish the type of collaboration as enabled by their web design. It may involve just piling up all contributed contents ('loose collaboration') or also working on a collectively evolving product ('tight collaboration'). Or in the terminology proposed by Dutton (2008): collaboration may range from co-contributing (2.0) to co-creation (3.0). The other basic parameter for communities of open content is their conditions of admission to the work process, which may range from fully open access to more restricted access.

The open invitation to contribute yields variable results. Although a large majority of contributions are usually valuable for the goals of the project, invariably disruptive and damaging contributions are coming in as well: they are off-topic, inappropriate, improper, offensive, and/or malicious, and so on. Obviously, the more a community leans towards full-blown co-creation ('open collaboration'), the more urgent the issue becomes, since disruptive edits may

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

[1] University of Groningen, Groningen, The Netherlands

actually endanger the integrity of the collective product as a whole.

In response, many open collaboration projects have taken to developing *anti-intrusion systems* that try to detect improper contributions and eliminate them before they can do any damage. Many types are in use, closely connected to the specific communities involved. Two dimensions are characteristic of such systems (de Laat 2015: pp. 175–176). On the one hand we may distinguish *who* is involved in them: managing editors employed by the project (who may hire additional workers), a selected elite of users, and/or all users without distinction. On the other hand we have to distinguish, for the personnel involved, *what* they are allowed to do about new contributions: check them, vote on them, and/or correct them. In general, the more massive the disruptions to a community, the more pressure builds towards enlarging the anti-intrusion work force (by mobilizing more ordinary users) and/or granting them more powers (powers of correction in particular).

A prime example of an open collaboration community plagued by damaging disruptions is Wikipedia, the co-creative encyclopedia with full write-access for all. Although all language versions suffer from vandalism, its English language version does so in particular. How large exactly is the phenomenon of vandalism in the English Wikipedia? Against a background of over 5 million entries, growing at a rate of 800 new ones a day, Wikipedia daily receives 90,000 fresh edits from human contributors. About 8.5 % of those may be estimated to be instances of vandalism. So every day Wikipedia has to deal with as many as 7500 malicious edits.[1] In response, various approaches have been tried out and tested; some have endured, others have perished. My focus is on the approach that has carried the day: a massive mobilization of Wikipedian volunteers to monitor and survey new edits around the clock. From administrators at the top to ordinary users at the bottom, all are asked to do their part; moreover, fully autonomous bots are enlisted as 'co-workers'. These mobilization efforts are facilitated and enhanced by the development of an array of anti-vandalism tools.

This system of surveillance, carried out by humans, bots, and tools, has been described before and analysed in view of the moral questions that it raises (de Laat 2015). It was found, to begin with, that, although all Wikipedians are invited to watch out for vandalism and revert any instance of the kind, the *stronger* tools in the counter-vandalism repertoire which allow faster search and correction are only distributed to trusted users. This policy has been adopted since the tools can do much damage. Moreover, I argued that these stronger tools may favour quantity over quality while checking edits and cause a loss of the required moral skills in relation to newcomers. In general, the system was found to operate in an invisible and opaque fashion, well hidden from sight to ordinary users. Besides these questionable issues there was one more issue that I brought to the fore. In order to facilitate the process of selecting edits for inspection, effectively *profiles* are being constructed and put to use. Some dimensions of those profiles appear to be problematic (de Laat 2015: pp. 181–182, section on 'profiling'). In the remainder of this article I analyse this claim in a more complete fashion—two pages cannot do justice to the complexity of the issues involved. So this article continues the discussion about profiling initiated earlier. Nevertheless it is intended to stand on its own; therefore all details necessary for the discussion will be reproduced below from the earlier publication.

The analysis proceeds as follows. The Wikipedian tools for edit selection and edit correction are extensively described; an important supporting element is the deployment of algorithms for calculating vandalism probabilities. After this exposition I give an overview of what is generally meant by profiling, and develop an account of how Wikipedia engages in profiling; a spectrum of increasing profiling for anti-vandalism purposes is distinguished. Subsequently I tackle the questions whether and to what extent these profiling practices are *effective* and *efficient,* as well as *morally justified*. In order to do so, I draw on two treatises by Frederick Schauer, an American philosopher of law.

His *Playing by the Rules* (1991) provides a framework to judge the (dis)advantages of the system of profiling as a specific system of rules. Thereafter, his *Profiles, Probabilities, and Stereotypes* (2003) provides a useful background to discuss complications resulting from the specific choice of profile features. On the one hand, particular dimensions may be 'overused' by human rule-enforcers, thus undermining profile efficacy. On the other hand, features may represent sensitive dimensions (such as race and religion) that may stir up social tensions—or create them in the first place. These general insights provide a lens to analyse and comment on the dimensions used in Wikipedian profiling.

Wikipedia: anti-vandalism tools[2]

Fighting vandalism basically consists of two stages. In the first stage ('selection') a new edit to the encyclopedia is selected for inspection; in the second stage ('inspection')

[1] All figures derived from https://stats.wikimedia.org/EN and https://en.wikipedia.org/wiki/Wikipedia:Vandalism_statistics.

[2] The two sections that follow—on tools and algorithms—are abstracted from de Laat (2015). They are the necessary building blocks to start the discussion proper about profiling.

the edit is actually inspected. If it is found to be obviously vandalistic it gets deleted (reversed); if it is found to be a bona fide edit, it is left intact. In reality, of course, borderline cases may turn up: the issue is not always so clear-cut. In such cases, a patroller may choose to act on these doubts by leaving a message on the editor's talk page, amending the edit involved, and the like.

Concerning the first stage of edit selection, new edits—which come in all the time—can be displayed on the screen in an ever-continuing list. Since inspecting all of them is impossible in view of the numbers involved, any patroller has to make some selection. It is precisely at this point in the process that several tools facilitate making this selection. First, the *type of entry* which has been edited may be selected. One composes a list of specific entries and watches only new edits to that selection of entries (in Wikipedia a so-called 'watch list' can be created for the purpose). Similarly, entries about living people can be watched closely. Secondly, one may focus on features of *content*: edits containing bad words, with massive blanking, either in part or as a whole, etc. Thirdly, *editor* characteristics may be focussed on: contributors who are anonymous (i.e., they have not registered, have no personal account), are new, have been warned, have been blacklisted before, etc. In the opposite vein one may choose to ignore edits made by certain types of contributors: administrators, bots, whitelisted users, and the like. Not unimportantly, to some extent filters can be combined and applied together; an obvious combination would be selecting anonymous contributions containing 'bad words'.

Subsequently, after inspecting the selected edit, the patroller may revert it if it is diagnosed as vandalistic. Such edit reversal can be supported by several buttons that allow performing instantly appropriate follow-up actions: leave a warning message on the talk page of the vandal, ask for administrator intervention against him/her, ask for the page to be 'protected' (i.e., categories of users are temporarily excluded from contributing, typically users who have just recently registered or not at all), and the like. Without these buttons, actions of this kind are cumbersome to perform.

These supportive functions for selecting and inspecting have, in various combinations, found their way into a range of concrete tools. The main ones are displayed in Fig. 1 (copied from de Laat 2015)—several of them will be discussed more fully below. For the moment let me, for illustrative purposes, just mention the #cvn-wp-en freenode channel. On this channel, IRC bots continuously broadcast fresh edits deemed suspicious. Moreover, the reason(s) for suspicion are specified as well: possible gibberish, large removal, blanking, etc. They obtain their colouring according to relevant editor characteristics: purple for a normal user, dark green for an anonymous user, red for a blacklisted user, and so on. So a multiple focus for selecting new edits can easily be practised.

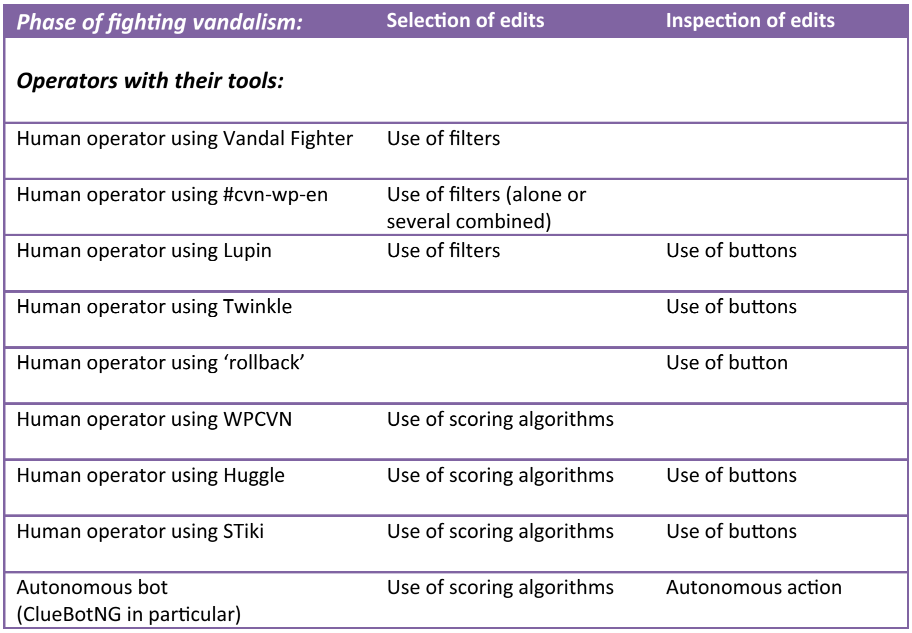
Wikipedia: algorithms

The most recent boost to fight vandalism has come from the development of computational approaches. Algorithms of the kind calculate the probabilities for each edit that it is actually vandalistic. Four varieties have been developed so far (Adler et al. 2011). As far as content is concerned, they may focus on language features (e.g., bad words, pronoun frequencies), or on language-independent textual features (e.g., use of capitals, changes to numerical content, deletion of text). A third type focusses on so-called metadata (e.g., time and place the edit was made, anonymous editor, warned editor), while a fourth and last type focusses on the editor's reputation as a trustworthy contributor, and on the text trust of the article involved (i.e., its reputation as it is revised by trustworthy editors).[3] All measures have something to say for them—although reputation sometimes has to be ruled out as being unreliable. Empirically, after a computer tournament with all approaches participating, it has been concluded that a combination of all four—if feasible—works best.

These algorithms have been incorporated as 'engines' in anti-vandalism tools. On the one hand, they figure in *'assisted editing'* tools like Huggle and STiki (Fig. 1). Let me describe the workings of both tools. As concerns STiki, the more sophisticated tool of the two, at its back-end new edits pulled from the Wikipedia servers are continuously fed to the engine. Edits are then classified by means of a specific method of machine learning: an alternating decision tree (ADTree). The most fitting values for the tree have been obtained before by training the model off-line on a reliable dataset of Wikipedian edits from the past (comprising both vandalistic and non-vandalistic edits); a dozen edit features of the third variety (metadata) were used in the analysis.[4] As an outcome of this supervised learning the classifier is incorporated into the STiki software and calculates the vandalism probabilities for incoming edits. Subsequently, suspect edits are passed from the back-end to the front-end and offered to human STiki operators in an ordered queue for inspection; patrollers have to process them from the top. Edits can be

[3] The difference between metadata and reputational measures—both at the metalevel beyond the edit itself—is just a matter of definition: metadata can be obtained immediately from edits as they appear on the Wikipedia server, while reputation is the outcome of—often complex—calculations that require data from the past.

[4] Currently, the outcomes of the neural network approach as employed by ClueBotNG (see below) can also be chosen as an alternative engine.



Phase of fighting vandalism:	**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 (alone or several combined)	
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 (ClueBotNG in particular)	Use of scoring algorithms	Autonomous action

Fig. 1 Anti-vandalism tools in Wikipedia and their affordances beyond the 'basic mode' of fighting vandalism (selection; cf. http://en.wikipedia.org/wiki/Wikipedia:Cleaning_up_vandalism/Tools)
Notes: '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 be usefully employed together (e.g., Lupin and Twinkle; WPCVN and Twinkle); WPCVN is out of order since January 2014. The tools of Vandal Fighter, #cvn-wp-en, Lupin, Twinkle, and WPCVN are available to all Wikipedians, the stronger tools of ollback, Huggle, and STiki require special permission.
Source: de Laat (2015)

accepted (classified as either innocent or pass; the latter option signals that the patroller is not quite sure) or reverted (either as vandalism, or as good-faith edit if no malicious intent seems to be present). Moreover, by means of several buttons comments and/or warnings can be instantly posted on the editor's talk page.

Huggle works in a similar fashion: incoming fresh edits are assigned a vandalism probability, based on simple manual scoring rules (in particular on the number of 'bad words'). The number of features taken into account is also a few dozen. These edits are subsequently offered to patrollers for selection in an ordered queue, with the higher probabilities on top. Unlike STiki, though, edits are shown with various colours which indicate suspicious features (such as editor has been warned, reported, or blocked). Patrollers may then be guided by these colours in their selection of edits from the queue for inspection—they are free to choose among them. Subsequent actions after edit inspection—acceptance, or reversion and posting a warning—proceed in a similar way again as in STiki.

On the other hand, these algorithms serve as engines for several *fully autonomous bots* (Fig. 1). These operate rather like the assisted editing tools just described, except that the operator in charge is not made of flesh and bones but of silicon. The decision to revert is made on the basis of vandalism probabilities; the ones above a certain threshold are reverted automatically and a message to that effect is posted on the vandal's talk page. At first calculations were based on manually written scoring rules (lists of bad words were the humble beginnings). Gradually, now, machine learning is taking over. The prime example of this approach is ClueBotNG. A Bayesian classifier has determined optimal vandalism weights for words and combinations of words in edits; these scores have been used as input for artificial neural network learning. About 300 edit features have been taken into account in the process. The model has been trained on a dataset of good and bad edits as classified by humans. Its output for any fresh edit is used as the edit's vandalism score. It is this trained bot which manages to check all edits coming in and reverts about one every minute. In terms of numbers, this bot is the top patroller of all time; it has reverted millions of edits since its inception in 2011.

Wikipedia: profiling

After this description of the whole array of counter-vandalism tools and bots in use, let me focus specifically on the first stage of patrolling: selection of new edits for closer inspection. My argument is that the forms of selection practised, from the most basic form up to the most sophisticated form, represent ever so many stages of sampling as steered by a *profile*. What do I mean by the term?

Originally profiling referred to constructing a bundle of personal characteristics meant to indicate the person or persons one is looking for. Criminal—or offender—profiling is the archetype: the police are searching for a criminal who fits a particular profile. So originally profiling was very *person-based*: the data on which the profile was based were personal data, the target was a specific individual. Gradually, however, the term profiling has acquired a much broader meaning, in particular regarding the target, the data collected, and the underlying techniques.

a. *Kind of target* It might seem that as a rule profiling targets specific individuals: persons who deviate from the norm that is involved. Officials are looking for a criminal, for persons who illegally crossed the border, for drivers who speeded through a red light, and the like. But often enough, a profile is not intended to catch individuals but to catch *acts of deviance*. Officials are on the lookout for instances of tax evasion, money laundering, drugs trafficking, contraband smuggling, or boarding a plane with explosives. So properly speaking not deviants are targeted but acts of deviance.[5]

b. *Kind of data collected* As a rule, nowadays not only data of a personal nature but *any* data are collected that have a connection with the act of deviance that is the target. In particular behavioural data and data about the particular transaction concerned are (also) employed for use in a profile. For the purposes of detecting income tax evasion, money laundering, drug trafficking, or smuggling, officials routinely collect indicators of behaviour deemed suspicious (Schauer 2003; Canhoto and Backhouse 2008; Zarski 2011).

c. *Underlying techniques* Profiling has turned into a sophisticated process of pattern recognition that uses large databases and employs techniques such as data matching and data mining. Its essence is discovery of knowledge; profiles are being constructed in an inductive fashion (Hildebrandt 2008). Anrig et al. (2008) give an overview of the basic techniques of data mining involved. Output for decision-making is delivered by either a black-box approach which is basically opaque (typically neural networks); or by a structured decision-making process that can be read and interpreted by humans—and explained to other humans too (typically classifiers and decision trees) (cf. also Canhoto and Backhouse 2008).

So taken together the conception of profiling has acquired a more general meaning that is useful across a range of situations. Steinbock nicely catches the connection with the criminal profiling of old: "Data mining's computerized sifting of personal characteristics and behaviours (sometimes called 'pattern matching') is a more thorough, regular, and extensive version of criminal profiling" (Steinbock 2005: p. 4). In a way, economists have always employed this broad conception of profiling when they define it as "the prediction of outcomes of interest conditional on observable covariates and the use of such predictions to make decisions regarding the members of a population" (Epple et al. 2006: p. F460). Another term they use is 'statistical discrimination'.

If we now turn our attention to Wikipedia again, the anti-vandalism tools described above can easily be interpreted as ever so many instances of profiling. First observe that these efforts are focussed on vandalistic editing, on catching malicious edits as soon as possible after having been contributed. The focus is definitely *not* on identifying and catching supposed vandals—I know of just one author pleading for such an emphasis (Kumar et al. 2015). Further, the kind of data being used in the tools described are as broad as possible, mainly behavioural (about the contributor: metadata, reputational data) and transactional (about the edit itself: language and textual features). Finally, the more sophisticated tools rely on generating a pattern by means of either structured decision-making (decision trees) or a black-box approach (neural networks). So the anti-vandalism tools exemplify the broader type of profiling.

If this profiling is done properly, Wikipedian vandalism fighting promises to yield more hits than are obtained by simple random sampling. Let me survey the various approaches from this angle. The Wikipedian patrollers who use no tools (as listed in Fig. 1) whatsoever are obviously not involved in profiling. They may just be looking at a screen full of new edits—and per force just take a random sample (since speed defies their checking all of them). Or, alternatively, they may take a special interest in specific entries that they want to keep free from vandalism. Accordingly they focus on fresh edits to these entries in particular—their sampling is 'subject-based'. Next consider patrollers who employ the less sophisticated tools from Fig. 1: tools such as the freenode channel or Vandal Fighter guide their selection process. As explained above, such tools allow filtering new edits along one or more dimensions. That is, said patrollers, based on their own personal experiences, reason—whether correctly or not—as to which dimensions promise the best catch, and decide to trawl accordingly. In other words, after careful reasoning

[5] For ease of exposition I only mention norm deviance that is guarded by state officials here. But of course in the private sector it has long been standing practice to use profiles in order to assess risks when serving customers. In this vein insurance companies estimate insurance risks and banks assess customer creditworthiness for lending purposes.

they decide to construct a specific profile and subsequently apply it to the fresh data. I henceforth denote this screening as ‘informal’ profiling (cf. also Schauer 2003: p. 173).

Patrolling in its more sophisticated form uses data mining approaches, be it performed by humans (assisted editing) or by bots. An elaborate profile is constructed, counting up to dozens of dimensions (Huggle, STiki) or hundreds of them (ClueBotNG). The profile directs the attention of patrollers in a forced manner: a queue of edits with the highest scores on top is presented to them, like a pile of cards to be dealt with. While using Huggle a human patroller may still choose from among the queue and jump in at any point, using STiki one no longer has a choice: one can only proceed by pronouncing one’s verdict on the edits in the queue (a batch of 5 edits), one after the other. One may switch queues in STiki (from the metadata to the ClueBotNG queue), but then it is the same story all over again—a new batch of 5 edits waits. Bot patrollers, finally, revert the highest probabilities on their own, leaving the other, lower probabilities to their human counterparts.

So in all, as concerns selecting fresh edits to Wikipedia for inspection we can distinguish a spectrum of increasing profiling: random or subject-based sampling (without using any profiles), informal profiling, formal profiling, up to automated profiling. It is this profiling that takes centre stage in this article. For one thing we ask: is this profiling profitable, does it bring the rewards that are usually associated with it? For another we ask: is this profiling approach towards edit selection justified? In particular, do any of the dimensions in use raise moral objections? If so, can these objections be met in a satisfactory fashion, or do such controversial dimensions have to be adapted or eliminated? It is these questions that I attempt to answer in the body of this article below.

Schauer: reasons for rules

Frederick Schauer has become famous for two books. His *Profiles, Probabilities, and Stereotypes* (2003) is a discussion of profiling in several contexts; I make use of it later on, while discussing the issues raised by the specific dimensions used in Wikipedian profiling. A decade earlier he published *Playing by the Rules* (1991) in which he discussed rules and rule-based decision-making in social life. Speaking in general he asks: what good are rules for regulating our behaviour? What can be said in their favour? In this section I give a brief summary of this discussion about rules. Then, in the subsequent section, I adapt his ‘reasons for rules’ in order to shed light on the rewards that Wikipedian profiling may bring. Note—oddly enough, from my point of view—that Schauer himself does not bring his 1991 arguments to bear on his discussion of profiling (as covered in Schauer 2003).

In chapter 7 of *Playing by the Rules* Schauer adduces some reasons for rules. From the outset, he resolutely pushes aside the argument from *fairness*: decision-making based on rules can only be *less* just than deciding each case on a particularistic basis (Schauer 1991: par. 7.1, p. 135 ff.). Rules force unlike cases to be treated alike, and may therefore deviate from an optimal decision that takes all particular circumstances into account. Some of his favourite examples are ‘Speed Limit 55’ for traffic and ‘No Dogs Allowed’ for restaurants. As to the former case, in some situations 75 Miles might be quite safe, while in other situations even 45 Miles is dangerous. Similarly, some dogs are capable of very civil behaviour in a restaurant, while other living creatures (such as snakes) may create great havoc. As Schauer phrases it: (simple) rules unavoidably suffer from *underinclusion* as well as *overinclusion* (Schauer 1991: pp. 31–34). Cases are underincluded, when they should really be included in the light of the relevant background justification but are not; cases are overincluded when they should be excluded in the light of this but are not.

The justification for rules therefore has to rest on *other* arguments that compensate for this sacrificing of fairness. He mentions several. To begin with, rules create *reliability/predictability* for those affected by the rule: rule-followers as well as rule-enforcers. They can plan their activities accordingly (Schauer 1991: par. 7.2, p. 137 ff.). This advantage only obtains if the promulgated rules are simple and widely known. ‘Speed Limit 55’, for example, makes life predictable for drivers, policemen, and judges alike.

Furthermore, rules promote more *efficient use of resources* by rule-enforcers (Schauer 1991: par. 7.3, p. 145 ff.). They do not need to immerse themselves in the precise details of each case, but can just apply the simple rule and decide accordingly. Rules allow them to sit back and relax almost completely. Obviously, their decision-making proceeds in a more efficient fashion. Concerning speeding (over 55 Miles per hour), for example, police officers and judges can now deal with it in an instant.

One more argument for rules is *risk-aversion* (Schauer 1991: par. 7.4, p. 149 ff.). Rule-enforcers who do not rely on rules but practice particularistic decision-making that takes all relevant factors into account, face a hard task. In the process they may quite well produce wildly erroneous decisions. Often, unfettered decision-makers produce a greater number of errors than those who (have to) respect a few simple rules. A system of law may want to avoid such risks, and introduce some carefully worded simple rules which curtail their discretion. The distrust experienced towards some sections of rule-enforcers may necessitate the partial revocation of the trust granted to them.

A final argument for rules is that they create *stability* in the system at hand (Schauer 1991: par. 7.5, p. 155 ff.). All the arguments just mentioned—reliability, efficiency, and risk-aversion—share a common focus on stability for stability's sake. Rules entrench the state of affairs that they have created in the first place. As long as that state of affairs benefits from being more or less permanent, the achieved stability is a desirable outcome. Necessarily, though, such stability is an impediment to change; it entrenches the status-quo. If change is on a society's agenda, the stability argument turns into an argument *against* having (simple) rules.

Wikipedia: substantive rules versus procedural rules

This eloquent defence of (simple) rule systems may produce an elegant approach to the issue of benefits of profiling in Wikipedia. Before proceeding, though, we should have a clear view of the sort of rules involved in profiling. The rules that Schauer discusses prominently in his 1991 book are rules that guide the decision-making of rule-enforcers towards issues such as driving too fast, taking your dog into the restaurant, etc. These are *substantive* rules. In profiling, a bundle of dimensions is taken together that subsequently guides the decisions as to which cases are to be inspected. The profile prescribes how to go about selecting people for inspection: these are *procedural* rules. Of course procedural rules are just the prelude to applying substantive rules later; after, say, singling out passengers at the airport (using a profile), their luggage gets screened (applying the luggage regulations in force).

If we now turn to a discussion of the benefits that profiling in Wikipedia may bring, it is immediately evident that the procedural nature of profiling rules changes the above discussion (from 1991) considerably. In particular, Schauer's first argument about the amount of justice produced by the introduction of rules has to be reinterpreted. Substantive decisions are taken with justice in view. Substantive rules may restructure such decision-making; as explained above, in order to achieve a series of clear benefits (such as predictability, efficiency, risk-aversion, stability), some justice is sacrificed. Sub-optimality is the price to pay. Our procedural rules involved in profiling, however, do not focus on justice; instead, I argue, they focus on *efficacy* straight away. This needs some explaining.

The discussion about introducing substantive rules starts from the default state of affairs that *all* cases produced have to be decided on. The comparative question is: in deciding on the cases brought forward, are we better off introducing (simple) rules or remaining without them and continue to judge them one by one? With our procedural rules, the baseline is of another nature. We start with an abundance of potential offenders, and realize that we have no means at our disposal to check and pass judgment on all of them; a selection of a kind *has* to be made. So it is here that profiling comes in. The comparative question is: in making a selection of cases to be inspected, are we better off introducing the tools of profiling or continuing to choose in random fashion? It is immediately clear that this comparative question has nothing to do with justice as such. Any one sample is not fairer than any other sample (since they are all candidates for inspection); the one can only be more *on target* than the other (i.e., bring more offenders to light than the other). It is a matter of *effectiveness*, not of justice.[6]

Wikipedia: benefits of profiling

If this shift in meaning is accepted, Wikipedian profiling turns out to be amazingly effective. Some indicators are the following. Sampling, whether at random or subject-based, only yields the average of vandalistic editing in general: about 8.5 % are 'hits'. By using proxies like anonymity or blanking as filters this rate is bound to increase. As soon as anti-vandalism algorithms are in operation, the situation becomes more complex since humans and bots become intertwined. The engines analyse *all* fresh edits and calculate the vandalism probabilities for *all* of them. Their output can be conceptualized as an ordered queue, with the highest probabilities on top. Subsequently, bots and humans—in that order—take their samples for treatment from this ordered pile. First the autonomous bots seize the highest probabilities on top and revert all of them—within a matter of minutes. The threshold level of estimated vandalism above which bots revert automatically has been set very high in order to avoid wrong decisions being made by the machine—it is calculated from a rate for so-called false positives that is deemed acceptable: 1 in 1000 reverts. After the bots, the humans armed with assisted editing tools are offered the chance to take their samples from the remaining pile. By definition, they receive the lower probabilities (below the threshold as defined) for inspection. What the bots may not and did not touch is allotted to them.

What about the hit rates involved? The bots obtain their batch of fresh edits (above the threshold as set) and just revert all of them; by definition, therefore, their hit rate is 100 %. Of course this does not imply their judgment is infallible; the lower we set the threshold, the more false

[6] This shift in meaning does not imply that all is well with profiling. As we shall see below, the pain with profiling lies elsewhere: with the choice of specific dimensions that make up the profile. Do any of them invite 'overuse' or unjustly discriminate against specific categories of people?

positives will be produced. As for humans, the rates they obtain with their samples are not fixed by definition but variable. As a follow-up on the inspection by the bot engines, they inspect the edits in their own human way. STiki scoreboards tell us, that their human patrollers commonly achieve a hit level between 15 % and 25 %. About one in four to one in six fresh edits offered for inspection gets classified as obvious vandalism. Note that this rate is bound to vary according to the level of vandalism at the time of patrolling and the amount of time humans actually spend patrolling. The more human beings are patrolling and the longer they work, the lower their rate becomes—the queue gets depleted.[7]

After this analysis of the effectiveness of Wikipedian profiling, we turn to the Schauerian argument of *reliability/predictability* for those affected by the rule. In the context of Wikipedia those affected by the rules of profiling are the contributors from all over the world as well as the patrollers who are constantly watching fresh edits. Profile-based patrolling changes the rules of the game. The nagging question of which edits are to be selected has been answered in an unequivocal manner: no longer those from a random—or, for that matter, a subject-based—sample but those corresponding to the profile. No more doubts for patrollers about where to look for vandalism; no more doubts, also, for potential vandals about the near impossibility of slipping through the net. As may be clear, the more we move from informal to formal profiling, the firmer the answers are, and hence the more reliability has been established. One could even argue, drawing this argument to its logical conclusion, that Wikipedian patrolling could benefit the most from complete transparency; all details of the profiling efforts should be available to the public at large. This would solidify the image of near-perfect patrolling for potential vandals and operate in pre-emptive fashion. In this light it is unfortunate that the anti-vandalism system in use remains opaque to ordinary users (as argued extensively in de Laat 2015).

The next Schauerian argument in favour of rules was that it enables rule-enforcers to use their resources in a more *efficient fashion*. This argument would seem to apply to the Wikipedian context in a straightforward sense. Patrollers, qua rule-enforcers, no longer have to develop their instincts about where vandalism may hide in the ever-continuing stream of fresh edits. Profiling tools make life easier for them. With intermediate tools like Vandal Fighter they may select one or two dimensions and apply them steadfastly. With assisted editing tools the whole business of pondering on and choosing a profile has even been taken over by the machine; patrollers just work the queue without having to bother about anything of the kind. Maximum efficiency in applying patrolling resources resides there. This argument indicates that in an ideal world each and every patroller should be able to resort to assisted editing tools; it is there that efficiency can be gained. In reality, though, the use of these tools is heavily regulated: only those who can prove their allegiance to the cause of Wikipedia may obtain permission to use them (for details, cf. de Laat 2015: pp. 180–181).[8] Out of fear of misuse and resulting damage, efficiency gets curtailed.

A further reason for rules was *risk-aversion*: in order to prevent erroneous decisions, the capricious use of discretion on the part of rule-enforcers becomes curtailed. For the Wikipedian context this argument is about patrollers running wild. They may think they are doing well in their selection of edits for inspection, while actually they achieve no more than a random hit score (of 8.5 %). So their activities are largely a waste of energy. As in the former argument, to which it is intimately linked, they would be well advised to turn to more sophisticated profiling tools—anything is better than plain intuition. Some nudging by Wikipedian 'authorities' would be helpful in this regard. But then again, as mentioned, fear of misuse of these strong tools largely prevents this.

Finally there is the argument from *stability*. Does it in any way apply to the case at hand of profiling within Wikipedia? For profiling in general stability—in the sense that the profiles in use are stable over time—is not always desirable. Consider for example the targeting of passengers at airports (for purposes of detecting drugs or explosives). This cannot just rely on static profiling (cf. Schneier 2005, 2012, 2015: pp. 136–140). All too often, potential miscreants test in experimental fashion which profiles are currently in use. Based on the results, they change 'personnel' in an effort to escape the controls. In such a game of cat and mouse, effective patrolling of passengers—if at all possible—can inherently only be dynamic.

That being said, I maintain that in the case of Wikipedia, profiling hardly needs to be dynamic. For one thing, the profiles in use contain many dimensions of edit content; these target vandalism rather precisely. There are no indications that the form and repertoire of vandalism changes

[7] Note that my particular definition of 'hit rate' implies the following. If bots were no longer to be allowed to revert autonomously (but just to calculate vandalism probabilities), the hit rate of humans would increase dramatically. If, on the other hand, the community were to decide to be more tolerant of mistakes made by a machine, the threshold level for vandalism reversal could be set lower. Accordingly, the autonomous bots would take over ever more anti-vandalism tasks from humans. Correspondingly, the human hit rate would decrease.

[8] There is a Counter Vandalism Unit (CVU) Academy in which potential vandal fighters may enrol and develop their capabilities under the tutorship of experienced patrollers (https://en.wikipedia.org/wiki/Wikipedia:Counter-Vandalism_Unit/Academy). But recruits have to work a heavy schedule to qualify.

much; correspondingly, the part of profiling that focusses on edit content is in no need of change. For another, profiles used contain metadata and reputational data (especially editor characteristics). Here, a game of cat and mouse has more room; but I have only sporadic evidence of vandals changing appearances in order to get through the controls.[9] So taken together this suggests that static profiling will do for the fight against vandalism in Wikipedia. Note that it is precisely this stability of the profiles in use that enables their efficacy, predictability, efficiency, and risk-aversion. If stability did not apply, the other arguments would become largely illusory.

Schauer: complications of profiling

The second book by Frederick Schauer that I want to bring to bear is *Profiles, Probabilities, and Stereotypes* of 2003. As the title indicates, it deals more specifically with the rules associated with profiling. Occupations treated are as diverse as tax officials (selecting taxpayers to undergo an audit), police officers (selecting and checking members of the public who look somehow suspicious), and airport officials (selecting passengers for drug screening or screening of their luggage). All of these officials (have to) take to profiling, whether of the more informal or the more formal kind as we shall see. The author sets out to explain the complications that the choice of specific dimensions for a profile may entail. For the sake of my argument in this article, two types of complications are important: (1) possible 'overuse' of dimension(s) (an issue of profile effectiveness) and (2) social sensibilities associated with specific dimension(s) (a social and moral issue).

Possible overuse Let me treat informal profiling first. Several dimensions are available to the officers involved which single out specific categories for inspection. Think of tax officers who check specific occupations (such as waiters, taxi-drivers, lawyers, and physicians) more closely than others (Schauer 2003: p. 163); airport officials looking for explosives who single out for inspection the luggage of younger Muslim men of Middle Eastern appearance (Schauer 2003: p. 181 ff.); airport officials looking for drugs who preferably select African-American women for a body search (Schauer 2003: p. 176 ff.); or police officers who preferably halt African-Americans for routine checks (Schauer 2003: p. 191 ff.). Obviously, the dimensions used should *make sense*; that is, be causally related to the offence involved. Then, using that filter yields more hits than random sampling alone. Such was the case with the first two examples just mentioned: tax officers and airport officials looking for explosives. Their actions were based on non-spurious correlations. However, with the other two examples mentioned this was *not* the case. The correlation between African-American origin (let alone only women of the kind) and drug traffic or street offence turns out to be non-existent (spurious). As a result, the efficacy of such a profile does not, on average, go beyond the level of random sampling. The reason why the officers involved acted so is rather obvious: they had either unthinkingly adopted some general stereotypes, or they were just outright racist (Schauer 2003: p. 177).

This efficacy argument becomes even more pronounced if profiling becomes *formal*. That is, a whole list of indicators of those likely to be offenders has been drafted, and in some manner guides the selection of whom is to be inspected. The Internal Revenue Service drafts an audit profile that creates a total score from over 50 features that correlate with the likelihood of mistakes or fraud in income tax returns (Schauer 2003: p. 162 ff.); customs officers employ up to 20 dimensions in profiling drug couriers (Schauer 2003: p. 169 ff.); and so on.[10] The whole point of these profiles is that, taken in their entirety, they square so strongly with the deviant behaviour that is targeted, that the officers involved in their application are well advised to stick to them and not stray from the path. Any improvisation that involves tweaking the profile, for whatever reason, undermines the effectiveness of the profile; fewer deviants are caught.

But the temptation does exist. Officers have a whole list of indicators at their disposal. Subsequently, they may choose to focus on some of them while they are immediately visible and therefore appear salient (Schauer 2003: p. 187). At the airport, officials ask *all* Muslim men and *all* men of Middle Eastern origin to step out of line to be searched. As a corollary, the other indicators suffer and much of the point of a profile gets lost. Schauer (2003) dubs this phenomenon the 'overuse' of some of the features of a profile.

This overuse—and attendant inefficacy of the profile—may be grave enough to think of measures to prevent it. As a matter of fact, as soon as socially sensitive features such as race and gender are involved, one may prohibit the use of such factors altogether: their use gets forbidden. The optimal profile is stripped of the indicators in question. Compensatory 'underuse'—or, I would say, more accurately non-use—gets realized. An interesting empirical

[9] In particular, the proxies of time of day and day of the week, location, and anonymity have constantly remained strong indicators of vandalism (to be discussed below).

[10] For an overview of past and present of aircraft passenger surveillance in general and American screening programmes like CAPPS, CAPPS II, and Secure Flight in particular, cf. Kite (2004) and Dummer (2005). A useful annotated bibliography with a section about passenger screening and surveillance is Tukdi (2007; slightly updated in 2014).

question then emerges. Does the stripped profile (non-use) turn out to be more effective after all than the original profile with several dimensions being overused in practice? Does it yield efficacy 'points' to curtail overuse that occurs so easily?

Usually though, there are other, more important reasons—not at all connected to profile efficacy—to mandate underuse of features: they are connected to social injustices that may increase precisely by their incorporation in profiling. This brings us to the second type of complications in profiling as discussed by Schauer (2003).

Social injustices

Dimensions involved in profiling, whether informal or formal, may touch upon socially controversial issues. One only has to think of factors like race, religion, ethnicity, nationality, or gender to realize the sensibilities involved. Take 'Driving while Black', a term used sarcastically to denote the experience of black people being harassed by excessive traffic controls (Schauer 2003: p. 191 ff.). Apart from the issue of efficacy, the drivers who are requested to stop feel harassed. Any driver stopped and frisked by the police feels harassed, but the point is that black drivers feel *more* harassed than others by the experience. This is so while they feel being discriminated against simply *because* they are black. As a result, the harassing experience does not fade away in time; instead, they feel hurt, their feelings of resentment and distrust towards the authorities increase. The essence of this phenomenon is rooted in the background that they *already* feel discriminated against; existing harms are magnified every time they are searched. This is called the 'expressive harm hypothesis' (Risse and Zeckhauser 2004).

The social tensions that such indicators may engender can be enough reason to call for abandoning them: they should no longer be used for targeting offenders. Other moral arguments sometimes strengthen this call. Take the above case of targeting black people. One may argue that a society should not be divided along racial lines. Anything that may prevent these lines from solidifying is to be done. Associated with this one may argue that even if Afro-American origin correlates with a higher rate of traffic offenses, such behaviour is also an outcome of age-long discrimination of black people in social life. *Because of* their being treated as second-rate citizens, black people engage in more unlawful behaviour such as speeding. A society should not only be wary of *continuing* this injustice—it should also *compensate* for the past effects of discrimination (Schauer 2003: p. 153).

Arguments of precisely this kind—in several combinations—have led to injunctions against using specific dimensions in profiling endeavours. Already since 1997, at least in the USA, while considered to be of a 'constitutionally suspect nature', the factors of race, religion, ethnicity, nationality, and gender have expressly been excluded from profiling—whether manual or automated—for purposes of luggage control at airports (recommended by a commission chaired by Al Gore).[11] In other contexts as well, mounting jurisprudence indicates that factors such as race and gender cannot be considered suitable indicators for profiling.

So note what is happening here: dimensions which are experienced as social sensibilities are expressly excluded from constructing profiles, irrespective of the fact whether they belong there or not for reasons of efficacy. Even if some dimension would score high after extensive and lengthy calculations on the data at hand and should by the very principles of profiling be included, this opportunity to gain efficacy is forfeited. Empirically, the comparative question becomes: what is lost in efficacy, is it a price worth paying for avoiding expressive harm and social discrimination?

After this survey of some of the complications associated with profiling in general we return to Wikipedia. The Schauerian questions to be pursued are (1) whether overuse of features involved in profiling is likely, and (2) whether some features happen to be sensitive issues in the online encyclopedic community. Figure 2 depicts these questions, and foreshadows the logic of the answers to be developed in subsequent sections.

Wikipedia: likely overuse?

A whole array of dimensions is available with which new edits can be filtered for inspection, either more informally or more formally (by means of a full profile). What about the issue of their efficacy? Important dimensions used are features of language and text (see above). Does the edit contain bad words, a string of exclamation marks, capitals only, or have chunks of text been deleted? Such features most likely correlate with vandalism. I had better put this more precisely: they do not indicate vandalism; they—almost invariably—*constitute* vandalism. They do not indicate where to look for vandalism; quite often they are the very thing itself.[12] So their contribution to an effective search is out of the question.

[11] "No profile should contain or be based on material of a constitutionally suspect nature—e.g., race, religion, national origin of U.S. citizens. The Commission recommends that the elements of a profiling system be developed (…) to ensure that selection is not impermissibly based on national origin, racial, ethnic, religious or gender characteristics" (White House Commission 1997: par. 3.19).

[12] Unless, of course, the edit involved requires the use of such features. Just consider the task of contributing to an entry about say pornography.

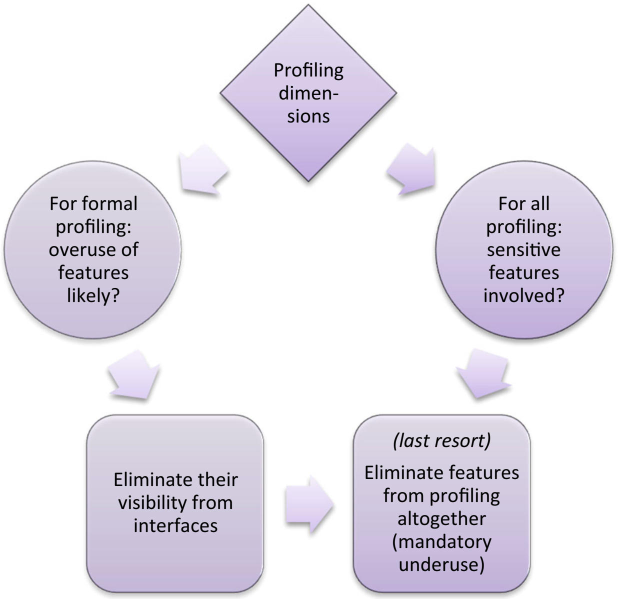

Profiling dimen-sions
For formal profiling: overuse of features likely?
For all profiling: sensitive features involved?
Eliminate their visibility from interfaces
(last resort)
Eliminate features from profiling altogether (mandatory underuse)

Fig. 2 Questions and actions pursued for dimensions used in Wikipedian profiling (based on Schauer 2003)

Next consider so-called meta-data. These are obviously (potential) indicators—not the thing itself. Does the editor have a newly created account, has he/she been warned or reported before, is he/she on a blacklist? It stands to reason that on average, correlations do occur; their use would seem to contribute potentially to efficacy. The same goes for being anonymous: unregistered users have consistently been found to behave in more vandalistic fashion.[13] Furthermore, measures of reputation may be helpful—the higher a user's reputation, the less likely that he/she is a vandal. Some less obvious features of edits are used as well: the time of day and day of the week they were made, and the country from which they originated. This has to do with the fact that vandalism is more prevalent at lunchtime on weekdays; employees and school pupils apparently try to chase boredom. Finally, edits from countries like the USA, Canada, and Australia are much more likely to be vandalistic than from elsewhere. So their use as warning signals also heightens efficacy.

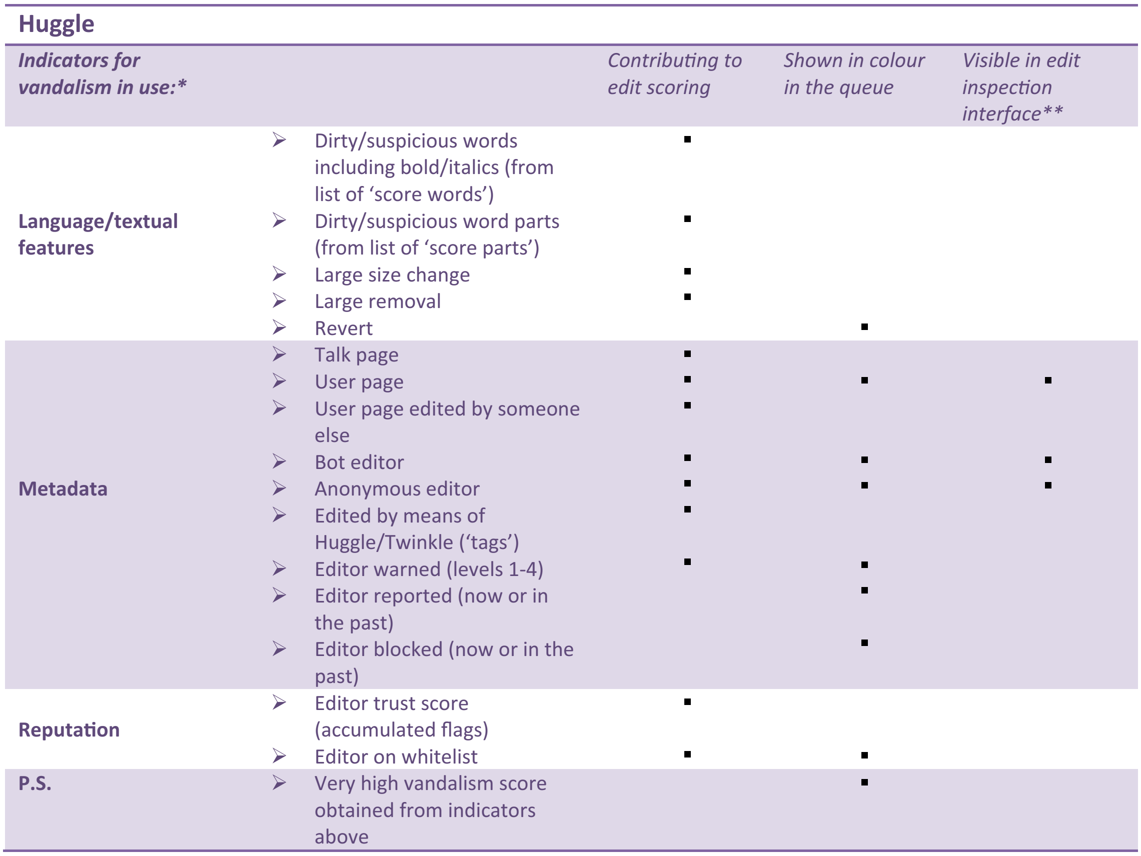
Now we have a full overview of the dimensions involved, we may turn to the first important question—as inspired by Schauer—about profiling dimensions: what about potential overuse of any of them? Is it possible that patrollers turn overzealous and unintentionally spoil the efficacy of profiling? In the case of informal profiling, patrollers use just one or two filters simultaneously; overuse does not seem to be an issue. At the other end of the spectrum we have the fully autonomous bots. Is there any danger that they indulge in overuse of the kind? The answer is (obviously) no. Bots (like ClueBotNG) just apply their algorithms for scoring new edits and subsequently revert those above the threshold-as-set. This is a fully automated procedure that cannot be tampered with by overzealous humans of a kind; overuse as defined is simply not an issue.

This is quite otherwise, however, for formal profiling with tools like Huggle and STiki. For both tools I see several possibilities for features to be overused by patrollers, thereby spoiling the optimum efficacy achievable by the profile embedded in those tools. In order to facilitate the exposition, I list the essential features of these tools in Figs. 3 and 4.

First consider Huggle (Fig. 3). It focusses patrollers' efforts in two ways: a profiling system that takes several—but by no means all—relevant features into account, and a colouring system that displays edits deemed suspect (as calculated by the profile) with a coloured square in front of them that indicates features such as editor is anonymous, has been warned (levels 1–4), reported, blocked, and the like. Notice, by the way, that in case several colours apply, only one takes dominance (colours do not mix). As can be seen from Fig. 3, edit characteristics relevant for anti-vandalism purposes have not been consistently parsed out between the one system and the other; some features are used in both. And that is where the problem of overuse starts to take shape. In order to explain this, let us first consider the—hypothetical but in my view ideal—situation that features have indeed been fully separated between the two systems. Patrollers obtain a queue (as ordered by the profile in use), and apply the features from the colouring system to subsequently filter from the queue and (hopefully) achieve even more hits than by just following the queue order. Since they employ *other* features than those incorporated in the profile, there is no issue of 'distorting' it. Applying skilful judgment they can only do better.

But in actual fact, relevant dimensions *do* overlap between the two systems (of profiling and of colouring). As a result, those dimensions are potentially used twice: a first time in the profile, a second time, possibly, by the patrollers using them as preferences for their filtering. If that happens, the dimension gets 'overused' and the profile's efficiency is affected. As can be seen from Fig. 3, the colouring system invites 'overselection' of two dimensions in particular: editor is *anonymous* (in grey), and editor has been *warned* (levels 1-4: in green, yellow, orange, and red).

[13] Some numbers about vandalism are the following. About 8.5 % of fresh edits constitute vandalism (7,500 a day). Most of such vandalistic edits concern 'obvious vandalism' or blanking. Almost all of them (!) have been contributed by anonymous editors (97 %). As a corollary, anons are estimated to be much more vandalism-prone than registered editors (by a large factor; as large as 20?!). Some 30 % of them have vandalized Wikipedian pages at least once. The figures just given are not very accurate—and rather variable at that. Nevertheless, the image of vandalism that emerges from them should by and large be correct.



Huggle

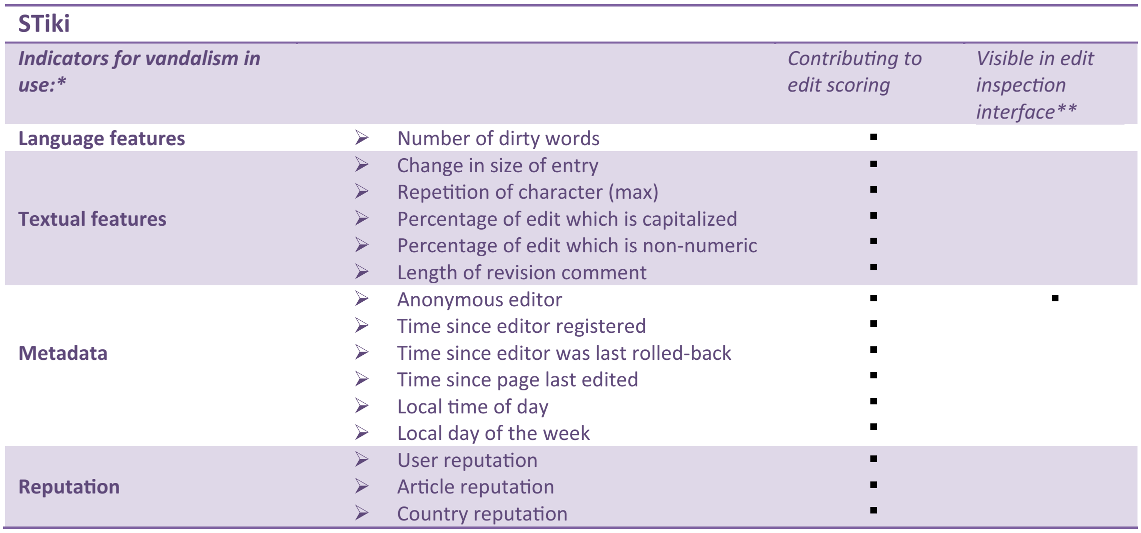
Indicators for vandalism in use:*		Contributing to edit scoring	Shown in colour in the queue	Visible in edit inspection interface**
Language/textual features	➢ Dirty/suspicious words including bold/italics (from list of 'score words')	▪		
	➢ Dirty/suspicious word parts (from list of 'score parts')	▪		
	➢ Large size change	▪		
	➢ Large removal	▪		
	➢ Revert		▪	
Metadata	➢ Talk page	▪		
	➢ User page	▪	▪	▪
	➢ User page edited by someone else	▪		
	➢ Bot editor	▪	▪	▪
	➢ Anonymous editor	▪	▪	▪
	➢ Edited by means of Huggle/Twinkle ('tags')	▪		
	➢ Editor warned (levels 1-4)	▪	▪	
	➢ Editor reported (now or in the past)		▪	
	➢ Editor blocked (now or in the past)		▪	
Reputation	➢ Editor trust score (accumulated flags)	▪		
	➢ Editor on whitelist	▪	▪	
P.S.	➢ Very high vandalism score obtained from indicators above		▪	

Fig. 3 Huggle: edit features that contribute to scoring edits for vandalism probability which determines queue order; those that are displayed in the queue in distinct colours indicating suspicion; and those that are visible in the interface for edit inspection
*As a matter of definition, metadata are immediately available, reputation measures require additional computation from past data. Notice that in the current configuration of Huggle the colouring of the metadata 'editor reported' and 'editor blocked' does not seem to be activated
**As far as metadata and reputation are concerned—language and textual features are visible by definition
Sources:
http://en.wikipedia.org/wiki/Wikipedia:Huggle;
https://en.wikipedia.org/wiki/Wikipedia:Huggle/Config;
https://github.com/huggle/huggle3-qt-lx/blob/master/huggle/wikiedit.cpp;
https://tools.wmflabs.org/huggle/docs/head/projectconfiguration_8hpp_source.html

Patrollers may be tempted to focus specifically on selecting such contributions from the queue—thereby overriding the queue order which is the result of a scoring algorithm that has *already* taken these features into account. This observation applies especially to the latter feature: colours indicating warning levels (four in all) abound all over the queue and constantly attract the attention of patrollers.[14] In addition, in the phase of edit inspection, patrollers may be tempted to give selected anonymous edits extra scrutiny and less benefit of the doubt precisely while they are visible as being anonymous (cf. Figure 3). For such edits vandalism inspection is more severe. 'Overinspection' joins overselection, which further jeopardizes profile efficiency.

As concerns STiki (Fig. 4), the possibilities for overuse appear to be more restricted—since the tool avoids the colouring system of Huggle that I just branded an invitation to overuse. Suspect edits are offered in a queue the patroller cannot escape from; they have to be judged one after the other. Options for filtering along personal preferences have simply been eliminated, thereby ruling out any potential overuse of features. But what about possible overuse in the subsequent phase of edit inspection? Most of the metadata and reputational indicators in Fig. 4 are not (or hardly) visible from the STiki interface, so any overuse seems to be ruled out. One of them is, however. Patrollers can actually observe whether an edit is made *anonymously* or not—and start inspecting them with extra scrutiny. Given that

[14] Note that the features of being a whitelisted user, being a bot, and 'user page is involved' also figure in both systems (Fig. 3). A special focus by patrollers in this case—of *ignoring* them—is not very likely though, since edits of the kind (almost) never show up in the queue in the first place—profiling has already effectively suppressed them before.



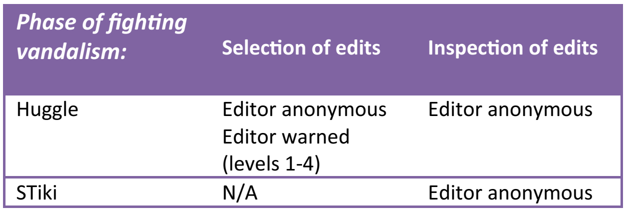
STiki			
*Indicators for vandalism in use:**		*Contributing to edit scoring*	*Visible in edit inspection interface***
Language features	➢ Number of dirty words	▪	
Textual features	➢ Change in size of entry	▪	
	➢ Repetition of character (max)	▪	
	➢ Percentage of edit which is capitalized	▪	
	➢ Percentage of edit which is non-numeric	▪	
	➢ Length of revision comment	▪	
Metadata	➢ Anonymous editor	▪	▪
	➢ Time since editor registered	▪	
	➢ Time since editor was last rolled-back	▪	
	➢ Time since page last edited	▪	
	➢ Local time of day	▪	
	➢ Local day of the week	▪	
Reputation	➢ User reputation	▪	
	➢ Article reputation	▪	
	➢ Country reputation	▪	

Fig. 4 STiki: edit features that contribute to scoring edits for vandalism probability (with the metadata approach) which determines queue order; and those that are visible in the interface for edit inspection

*As a matter of definition, metadata are immediately available, reputation measures require additional computation from past data. Notice that I consider revision comments to be part and parcel of an edit as a whole since they usually adduce reasons for an edit (they are not to be interpreted as being at the meta level). Note also that the dimensions of 'user reputation' and 'local day of the week' no longer figure in the most recent version (as of July 2013) of the classifying ADTree

**As far as metadata and reputation are concerned—language and textual features are visible by definition

Sources:
https://en.wikipedia.org/wiki/Wikipedia:STiki;
https://github.com/westand/STiki/blob/master/api/stiki_api_readme.txt;
https://github.com/westand/STiki/blob/master/learn_adtree/ex_model.txt;
https://github.com/westand/STiki/blob/master/learn_adtree/adtree_model.java



Phase of fighting vandalism:	Selection of edits	Inspection of edits
Huggle	Editor anonymous Editor warned (levels 1-4)	Editor anonymous
STiki	N/A	Editor anonymous

Fig. 5 Huggle and STiki: metadata and reputation features that can possibly be overused by patrollers in the phases of edit selection and edit inspection (derived from Figs. 3 and 4 above)

anonymity is a well-known alarm bell for potential vandalism, the possibility of overinspection is a realistic one.

It can be concluded that formal profiling may give rise to overuse of several features, in particular of editors being anonymous and/or being recently warned on their talk page—as summarized in Fig. 5 for the Huggle and STiki tools. What can be done to prevent this overuse from occurring at all? For one thing, one could eliminate the actual possibilities that *enable* overuse as offered by the interface design of the tools involved. With Huggle, the colouring system is just an invitation to overuse as far as features are incorporated that chances are will be used a *second time*. Why not sanitize that system by excluding all features from it that have already figured in the preceding process of profiling? In particular, why not omit coloured squares indicating editor status and warning levels?[15] Or better still, why not optimize the profiling process by including all features that are deemed relevant and dispense with the colouring system altogether? Instead of a halfway solution with both profiling and filtering, a system of full profiling gets installed. In addition to the above suggestions Huggle is to offer edits for actual inspection without revealing any details about editor status: only the contents themselves (comments included) are to be made available for inspection. As far as STiki is concerned, it does not suffer from two interfering logics for the selection of edits. Therefore the only recommendation to their developers is that they refrain from displaying editor status (registered account or not) on the patroller's interface.

This possible remedy against overuse, of eliminating the visibilities in question, is depicted graphically in Fig. 2 (left-hand side). Notice that in his exposition Schauer did not mention this option of disabling the visibilities that may invite overuse. Although similar options are imaginable for, say, screening at airports, the officials involved

[15] Notice that Huggle also employs a colour to indicate that the outcome of algorithmic scoring is high (Fig. 3). However, obtaining a high score automatically leads to ending up high in the queue; adding a colour to it will hardly change its status of urgency. Therefore overselection does not appear to be an issue here. Nevertheless, the signalling is superfluous and may just as well be omitted.

apparently insist on continuing their face-to-face screening of passengers standing in line.

Observe finally, that the remedy usually applied against overuse is mandatory underuse: *eliminate* the features from the profiling efforts (incorporated in Fig. 2 as the measure of 'last resort'). In our case: eliminate anonymity and warning levels from the algorithms that calculate vandalism probabilities. This would be a drastic measure since it ties the hands of the profilers and potentially cripples the profiling effort. Given that—as just argued—overuse can effectively be undercut by redesigning the options to act in that fashion while leaving profiling intact, for the moment this drastic measure does not seem to be necessary. Further below, though, I shall argue that one of these dimensions, anonymity, is controversial from a societal point of view; the only remedy for such sensitivities is precisely their elimination from the algorithms in use.

Wikipedia: controversial categories?

Let us return to the full spectrum of dimensions being used in Wikipedian profiling, whether it is done more informally (filters), more formally (assisted editing tools), or automatically (bots). Are any of them associated with societal sensitivities of a kind, sensitivities that one may not want to touch by expressly forbidding their use in profiling? Are there any dimensions comparable to race, religion, and gender, features that stir up so many emotions in the societal debate? Actually, many features in use look harmless enough. Whether editors are new, warned and/or reported, recently reverted, or have a low reputation—all of these constitute ever so reasonable warning signs to which no one can reasonably object. Not any kind of discrimination, whether already existing or newly created, seems to be involved.

Nevertheless, a few dimensions remain that merit closer attention. Worries might spring from targeting employees and schoolkids in their *lunchtime* (as part of the STiki tool). Is it justifiable to treat them as a special group that needs to be approached with suspicion? None of the arguments used to ban factors like race or religion seem to apply here. The group can hardly be called a proper group: it is just a snapshot in time, its 'members' do not even recognize each other as a 'group'. Correspondingly, no lines of social discrimination are created let alone solidified, no harm from the past becomes magnified or needs to be compensated. Therefore I see no reasons to argue for banning this 'lunchtime dimension' from STiki-profiling.

Another worry is constituted by targeting the few *countries* that are considerably more vandalism-prone than others (as part of the STiki tool again). Can it be justified to approach contributors from the USA, Canada, and Australia with suspicion? Aren't we in danger of stirring up nationalistic tensions? After all, the nationals involved have considerable more 'groupness' than in the foregoing observation. Let me first point out that no nationalistic lines of division seem to be present in the Wikipedian community; therefore no such lines can be hardened, no nationalistic past is in danger of being magnified, let alone in need of compensation. The only argument could be that dividing lines of the kind tend to be *created* by such profiling-by-country in the first place; nationalistic controversies might be ignited by it. This constitutes some kind of argument—but I do not consider this dim possibility strong enough to warrant banning the country-dimension and correspondingly sacrifice some profiling efficacy.[16]

More serious, though, is the targeting of contributors who choose to remain *anonymous*. Unregistered users are consistently targeted all along the profiling spectrum, whether the patrolling uses informal or formal profiles. With all of them, anonymity is a warning signal. I would argue that this special attention is fraught with danger since anons *already* constitute a controversial group within the Wikipedian community.

As a matter of fact, they are much more vandalism-prone than Wikipedians who have registered and operate from an account (see note 13 above). This undisputed fact has created a lot of animosity against them. Many members argue as a result that registration should become obligatory: the time of anonymity has passed. In opposition to this it is pointed out that some contributors are just passing by and do not care to take the trouble to register; an option that should remain available. Moreover—and more importantly—it is argued that, as a matter of principle, the possibility of contributing in an anonymous fashion may be of vital importance for those who are involved in social, political, or ideological controversies. For them, the only way to continue the discussion in Wikipedia without endangering themselves may be in the guise of anonymity.[17] For them, registering or not is not a matter of choice—circumstances force them to operate as an anon. Finally, the argument goes, while they are so many, IP-accounts taken together actually contribute a lot to the encyclopedia (even if each individually just offers one or two edits); it would be a pity to take the risk of losing their contributions.

[16] Anyway, as soon as the dimension of anonymity no longer features in profiling (as I propose below), the issue would evaporate while only anons are currently targeted by nationality.

[17] Ironically, though, contributing from a registered account (after choosing a suitable pseudonym) keeps a lot more personal data away from outside prying eyes; on the Wikipedia servers unregistered users appear with their IP-accounts, which reveal features like location, time, and more.

It is against this background that the targeting of anons should be judged. Is the dimension of anonymity by any chance to be eliminated from all profiling efforts used in Wikipedia? My answer is in the affirmative. Not so much for tangible harm done to anonymous contributors when they are part and parcel of profiling schemes. After all, normally they do not experience any harm when their edits are selected and inspected as a result of anon-powered profiling; they will not even notice that they were surveilled since no digital traces remain of the patrolling. Of course when the patrolling *does* catch a vandal, things are otherwise: the perpetrator *is* affected and could notice as a consequence—but then it is his/her just deserts. The only imaginable harm is that patrollers become over focussed on anons and indulge in what I called above 'overinspection' of such edits and wrongly classify them as vandalism—just to be on the safe side; some innocent anons will be 'mistreated' as vandals. As a consequence, they might never contribute to Wikipedia again. In addition, the patrollers concerned might become ever more obsessed with chasing anons, creating a vicious circle that leaves nonregistered contributors little leeway. Nevertheless, I estimate this harm to be small. At any rate, the harm involved would seem to be small in comparison with the harassment of racial profiling—let alone that an 'expressive harm hypothesis' applies.

Instead, my main argument for the ban is a decidedly moral one. From the very beginning the Wikipedian community has operated on the basis of a 'social contract' that makes no distinction between anons and non-anons—all are citizens of equal stature. Fierce discussions over the years have clarified that the community remains firmly committed to this principle while contributors may have good reasons to choose the cloak of anonymity (as explained above). Given this pledge you cannot just proceed and treat anonymous contributors with special scrutiny as compared to non-anonymous ones. It will not do to proclaim that all citizens are equal, but meanwhile treat unregistered citizens as less equal than registered ones. As a final contract-related argument, while some anons arguably misbehave and misuse the write-access granted to them, that does not justify placing *all* anons in undifferentiated fashion under close algorithmic surveillance.

In sum, the express profiling of anons turns the anonymity dimension from an access condition into a social distinction; the Wikipedian community should refrain from institutionalizing such a line of division. Notice that I argue, in effect, that the Wikipedian community has only two choices: either accept anons as full citizens or not; but there is no morally defensible social contract in between.

Therefore I argue that the anonymous-dimension should be banned from all profiling efforts (Fig. 2). Its underuse gets mandated, in all possible ways. First and foremost, its use as a scoring dimension in the algorithms involved in assisted editing should be discontinued (Huggle, STiki); the same goes for neural network enabled learning (ClueBotNG in particular): the anon-condition should simply not be specified as one the characteristics of edits. Furthermore, this condition preferably should disappear from *all* interfaces involved in patrolling in general. No 'recent IP-edits' button on the standard interface, no special colour for anon-edits that allows filtering in #cvn-wp-en or Vandal Fighter, no grey indicator for them in the Huggle queue, no rendering of anonymity in the edit inspection interfaces of either Huggle or STiki—the anonymity condition should be made to disappear, effectively enabling all patrollers to wear the veil of ignorance in this regard. The (small) price to pay for social stability within the Wikipedian community is a slight decrease in profiling efficacy.[18] Notice, finally, that arguably (as advocated elsewhere: de Laat 2015) Wikipedia should make itself more accountable to the outside world. As part of this, total *transparency* of surveillance efforts is indicated. Seen in this light, underusing (banning) the anonymity dimension is of even greater importance. Targeting IP-accounts in the shadows is hardly tolerable already; targeting them in the limelight would broadcast the wrong signal completely.

Conclusions

The foregoing Schauerian exercise in justification of profiling as used in Wikipedia for the selection of fresh edits for inspection has yielded the following conclusions and suggestions. On the whole, the use of profiling tools has been found to be a blessing. Especially as profiling develops from more informal (filtering tools) to more formal (assisted editing tools) to autonomous bots in operation, the following benefits are realized. The effectiveness of the selection increases dramatically; patrollers may use their time and energy in more efficient fashion since the tools involved facilitate or take over the construction of profiles from them; assisted editing tools in particular contribute to curtailing the discretion of potentially erratic patrollers and channel them toward using preformatted profiles (risk-aversion); the stability of profiling rules—which seems appropriate to Wikipedian vandalism dynamics—serves as the foundation for the foregoing benefits.

[18] I have no hard data to accurately estimate this loss in efficacy. Nevertheless, since the number of indicators used is actually quite large (for the STiki metadata queue more than 10, for the ClueBotNG queue about 300), it is reasonable to estimate that the loss is small.

Some suggestions for change can also be derived from Schauer's reasoning about rules. Profiles introduce predictability as to which edits will be selected for inspection. This strongly suggests publishing all details of profiles in use, in an effort to create the image of near-perfect vandalism detection. As a result, potential vandals might be convinced on beforehand that their games can only backfire—and go elsewhere. Moreover, all benefits just mentioned become increasingly realized the more formal profiling becomes; assisted editing tools are simply superior to the other ones. This suggests strongly that these tools are distributed as widely as possible. As yet, however, fear of misuse by malicious patrollers has prevented this. Less than a thousand Wikipedians—who qualified as trustworthy—effectively have access to them.

A more detailed examination of the dimensions used in profiling yielded more severe points of criticism. They are related to the issues of overuse of specific features by human profilers and to the social sensibilities that are associated with some of them, necessitating their mandatory 'underuse'. Figure 2 depicts the questions that have been pursued.

While overuse is not an issue concerning either informal profiling or bot-operated profiling, it does turn out to be an issue concerning assisted editing tools. Huggle, in particular, invites overuse of the dimensions of anonymity and being warned before. Patrollers may disproportionally focus on selecting such edits (special colours alerting them to the feature); moreover, they may—unconsciously or not—subsequently proceed to inspect anonymous ones with heightened scrutiny. With STiki, overuse is also likely, though to a lesser extent: anons may get a more severe check because their edits are visible as such. Such overuses only serve to disturb the fine-tuned efficacy of the underlying algorithms.

Fortunately enough, a remedy seems readily available: the design of the corresponding interfaces is to be adjusted. Invisibility is the motto: anything that may invite patrollers to stray from the optimized profile (whether a small profile as in Huggle or a large profile as in STiki) is to be left out. For Huggle, the colouring alert system should either be trimmed down in order to avoid overlap or disappear altogether—in the latter case any and all relevant information is to be incorporated in the algorithm that orders the edit queue. Furthermore, for both Huggle and STiki, editor status (with or without an IP-account) is to be deleted from the interface that displays the actual edit for inspection.

Finally, do any of the features used in Wikipedian profiling as a whole (from informal to formal to bot-operated profiling) touch upon social sensibilities comparable to the use of race and religion in profiling elsewhere? Several features appear to be candidates for closer inspection of the kind. Targeting Wikipedians who contribute at lunchtime on weekdays, or targeting Wikipedians from the USA, Canada, and Australia merit critical attention—but appear to be rather harmless since they hardly incite one group against another. Targeting those who have not registered, the anons, however, is a different issue. Anons *already* constitute a contested category since, on average, they are clearly more vandalistic than registered users. Correspondingly, many Wikipedians insist that all contributors to Wikipedia register first and anonymity gets ruled out. I defend the position that Wikipedia is founded on a social contract that considers both registered and non-registered contributors as equal citizens. It will not do to break that contract while a small subset misbehaves and put the category of anons as a whole under special surveillance.

It seems imperative, therefore, to ban the anonymity feature from all filtering and profiling efforts altogether. The remedy has two components. For one thing, the feature should disappear from all interfaces which display suspect edits to patrollers (this *generalizes* the remedy for overuse mentioned above, for Huggle & STiki in particular, to *all* patrolling tools). For another, the anonymity dimension should no longer figure in algorithms which calculate vandalism probabilities. Anonymity is simply not a relevant dimension any longer. Both patrollers *and* programmers producing the algorithms involved henceforth wear a veil of ignorance as far as editor status is concerned. Its mandated underuse—or as argued: non-use—is the price to pay for social stability in the Wikipedian community.

Obviously, the requirement of banning the dimension of editor status (registered or not registered) would also have wider implications for any other vandalism detection system to be used by Wikipedians in the future. Let me mention two of these in particular. The ORES web service (under development) intends to provide vandalism edit scores on demand for any and all language versions of Wikipedia.[19] Furthermore, the so-called 'managed wiki' proposal essentially sorts edits (on beforehand) on the basis of anonymity and editor reputation (Wöhner et al. 2015). In my view, the software for both systems would have to be 'sanitized'.

I argue for elimination of the sensitive feature of anonymity. An intriguing challenge concerning this proposal may be mounted from the nascent field of 'discrimination-free' modelling (the sequel is based on Calders and Zliobaite 2013; Kamiran and Zliobaite 2013). Their practitioners argue that datasets may suffer from various defects (like incorrect labelling, sampling bias, and/or incomplete data); as a result training models (as used in data mining) by means of them produces biased output. In

[19] For more details, cf. https://meta.wikimedia.org/wiki/Objective_Revision_Evaluation_Service.

order to correct for the bias involved, a first obvious solution would seem to be to eliminate the sensitive dimension (in line with my suggestion above). However, there may be hidden correlations with that dimension in the datasets in use, which would allow discriminatory practices to continue ('redlining', 'masking'). One could go on and eliminate the correlated attributes as well, but every step of elimination usually eliminates some valuable information as well, thereby reducing the accuracy of predictions from modelling. A second obvious solution is to train separate models for each of the sensitive groups. However, this is bound to end up in a reversal of the original bias ('positive discrimination').

In order to find a way out of this dilemma attention has recently turned away from eliminating the sensitive dimension involved. Instead, the very models and their data sets for training are being reconsidered. How to train models in view of obtaining unbiased results (cf. Kamiran and Zliobaite 2013; Kamiran et al. 2013, Hajian and Domingo-Ferrer 2013)? In the pre-processing stage one may change the set of training data involved: locally 'massaging' the data in such a way that borderline cases are relabelled, and/or local 'preferential sampling' that deletes and/or duplicates training instances are options under consideration. In the processing stage one may take to developing models under non-discrimination constraints. In the post-processing phase, finally, one may try and suitably alter the classification rules obtained.

Some of these arguments may well apply to Wikipedia. The data sets of vetted edits used for the training of anti-vandalism tools may suffer from oversampling of anonymous edits (while patrollers may indulge in overselection) and carry incorrect vandalism labels (while patrollers may lean towards overinspection). As a result the classifiers and neural networks presently being used produce vandalism scores with a bias against anons. Just eliminating that dimension as I suggested above might conceivably not solve the problem of bias while running up against hidden correlations and/or deleting valuable information. Should the attention shift instead to modelling under anti-discriminatory constraints? A serious complication is that although part of the correlation between anonymity and vandalism is the result of bias against anons, part of it is real. One has to separate the two effects and choose one's constraints accordingly, before one can even begin the discrimination-free modelling. This complex question appears to open up a wholly new line of further inquiry.

Acknowledgments Thanks are due to Adam Shorland, Andrew West, and Chris Breneman—the developers of Huggle, STiki, and ClueBotNG, respectively—for their incisive comments on my analysis of their tools. Comments from various participants at the *Computer Ethics: Philosophical Explorations (CEPE)* conference (University of Delaware, Newark, USA) in July 2015 have also been helpful in strengthening the manuscript. The same applies, finally, to two reviewers from this journal.

References

All websites in this article have last been accessed on February 10, 2016

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.

Anrig, B., Browne, W., & Gasson, M. (2008). The role of algorithms in profiling. In M. Hildebrandt & S. Gutwirth (Eds.), *Profiling the European citizen, cross-disciplinary perspectives* (pp. 65–87). Berlin: Springer.

Calders, T., & Zliobaite, I. (2013). Why unbiased computational processes can lead to discriminative decision procedures. In B. Custers, T. Calders, B. Schermer, & T. Zarsky (Eds.), *Discrimination and Privacy in the Information Society: Data mining and profiling in large databases* (pp. 43–57). Berlin: Springer.

Canhoto, A., & Backhouse, J. (2008). General description of the process of behavioural profiling. In M. Hildebrandt & S. Gutwirth (Eds.), *Profiling the European citizen, cross-disciplinary perspectives* (pp. 47–63). Berlin: Springer.

de Laat, P. B. (2012). 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, December 5–7, 2012; Proceedings*, *LNCS 7710* (pp. 534–557). Berlin: Springer.

de Laat, P. B. (2015). The use of software tools and autonomous bots against vandalism: Eroding Wikipedia's moral order? *Ethics and Information Technology, 17*(3), 175–188.

Dummer, S. W. (2005). Secure flight and dataveillance, a new type of civil liberties erosion: Stripping your rights when you don't even know it (Comment). *Mississippi Law Journal, 75*(2), 583–618.

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

Epple, D., Romano, R., Sinan Sarpça, & Sieg, H. (2006) Profiling in bargaining over college tuition. *The Economic Journal, 116*(515), F459–F479.

Hajian, S., & Domingo-Ferrer, J. (2013). Direct and indirect discrimination prevention methods. In B. Custers, T. Calders, B. Schermer, & T. Zarsky (Eds.), *Discrimination and Privacy in the Information Society: Data mining and profiling in large databases* (pp. 241–254). Berlin: Springer.

Hildebrandt, M. (2008). Defining profiling: A new type of knowledge? In M. Hildebrandt & S. Gutwirth (Eds.), *Profiling the European citizen, cross-disciplinary perspectives* (pp. 17–45). Berlin: Springer.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2013). Techniques for discrimination-free predictive models. In B. Custers, T. Calders, B. Schermer, & T. Zarsky (Eds.), *Discrimination and Privacy in* *the Information Society: Data mining and profiling in large databases* (pp. 223–239). Berlin: Springer.

Kamiran, F., & Zliobaite, I. (2013). Explainable and non-explainable discrimination in classification. In B. Custers, T. Calders, B. Schermer, & T. Zarsky (Eds.), *Discrimination and Privacy in the Information Society: Data mining and profiling in large databases* (pp. 155–170). Berlin: Springer.

Kite, L. A. (2004). Red flagging civil liberties and due process rights of airline passengers: Will a redesigned CAPPS II system meet the constitutional challenge? *Washington and Lee Law Review, 61*(3), 1385–1436.

Kumar, S, Spezzano, F., & Subrahmanian, V. S. (2015) VEWS: A Wikipedia vandal early warning system. Obtained from http://arxiv.org/abs/1507.01272.

Risse, M., & Zeckhauser, R. (2004). Racial profiling. *Philosophy & Public Affairs, 32*(2), 131–170.

Schauer, F. (1991). *Playing by the rules: A philosophical examination of rule-based decision-making in law and life*. Oxford: Clarendon Press.

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

Schneier, B. (2005). Profiling. On his personal blog called Schneier on Security. Obtained from https://www.schneier.com/blog/archives/2005/07/profiling.html.

Schneier, B. (2012). The trouble with airport profiling. On his personal blog called Schneier on Security. Obtained from https://www.schneier.com/essays/archives/2012/05/the_trouble_with_air.html

Schneier, B. (2015). *Data and goliath: The hidden battles to collect your data and control your world*. New York, London: W.W. Norton.

Steinbock, D. J. (2005). data matching, data mining, and due process. *Georgia Law Review, 40*(1), 3–84.

Tukdi, I. (2007). Transatlantic turbulence: The European Union and United States Debate Over Passenger Data. Globalex, Hauser Global Law School Program, NYU School of Law. Obtained from http://www.nyulawglobal.org/globalex/Passenger_Data_US_EU.html

White House Commission on Aviation Safety and Security (1997) *Final Report to President Clinton*. Washington D.C.: The White House. Obtained from http://fas.org/irp/threat/212fin~1.html

Wöhner, T., Köhler, D. W. I. S., & Peters, R. (2015). Managed Wikis. *Business & Information Systems Engineering, 57*(3), 155–166.

Zarski, T. Z. (2011). Governmental data mining and its alternatives. *Penn State Law Review, 116*(2), 285–330.