Trolls, bans and reverts: Simulating Wikipedia

Valentin Lageard, Cédric Paternotte

Trolls, bans and reverts: Simulating Wikipedia

Valentin Lageard, Cédric Paternotte*

November 2018 - Penultimate draft, forthcoming in *Synthese*

Abstract. The surprisingly high reliability of Wikipedia has often been seen as a beneficial effect of the aggregation of diverse contributors, or as an instance of the wisdom of crowds phenomenon; additional factors such as elite contributors, Wikipedia's policy or its administration have also been mentioned. We adjudicate between such explanations by modelling and simulating the evolution of a Wikipedia entry. The main threat to Wikipedia's reliability, namely the presence of epistemically disruptive agents such as disinformers and trolls, turns out to be offset only by a combination of factors: Wikipedia's administration and the possibility to instantly revert entries, both of which are insufficient when considered in isolation. Our results suggest that the reliability of Wikipedia should receive a pluralist explanation, involving factors of different kinds.

1. Introduction

When Wikipedia appeared years ago, it was met with scepticism. How could an army of contributors, many of whom appear to have no special expertise, possibly end up producing any interesting, let alone correct content about topics, including scientific ones? How could knowledge collectively arise from the aggregation of amateurs or ignoramuses? Even if so, how could it possibly rival that of traditional encyclopedias? To the surprise of most, however, Wikipedia turned out to be quite reliable as a source of information. Although it is debatable whether this epistemic feat surpasses, rivals or comes close to that of traditional encyclopedias, hardly anyone could have confidently predicted it.

Why is Wikipedia a reasonably reliable source of knowledge? Tentative explanations abound. Maybe knowledge naturally emerges from the aggregation of ignorant individuals; or maybe enough of them are close enough to expertise regarding the topics they favour. Maybe Wikipedia is run by a close-knit community of elite administrators who strive for reliability and prevent abuse. Maybe the contributors mutually check and police each other. Maybe Wikipedia's technical features, such as history checks, discussion forums and votes are crucial in maintaining the overall quality of most of its pages.

The problem of explaining the reliability of Wikipedia goes beyond mere interest for this specific epistemic entity: it concerns social epistemology as a whole. Although the study of collective epistemic enter-

* The simulations were programmed and run by Valentin Lageard.

prises is not new, it has mainly focused on cases such as science and the social characteristics that make it epistemically successful. Such works typically assume that science stems from agents who aim for epistemic success or have motivations that happen to further epistemic success as a by-product. By contrast, the Wikipedia community contains a significant proportion of contributors with negative epistemic aims or have non-epistemic motivations that hinder epistemic ones; such contributors include disinformers and trolls.

Accordingly, our aim in this paper is to suggest explanations as to why Wikipedia is reliable despite the non-negligible proportion of contributors that oppose its epistemic aims.

We will be developing and employing computer simulations. This choice stems from the high number both of characteristics of Wikipedia that are possibly relevant to its reliability, and of hypotheses regarding the fundamental mechanisms that warrant such reliability. The simulation allows us to explore a number of possible scenarios and to isolate the specific roles of distinct factors. It also allows us to suggest a *pluralist* explanation – namely, the claim that it is a combination of factors that is integral to Wikipedia's epistemic reliability – which would have been more difficult to identify and defend otherwise. As we will see, this reliability does not boil down to any isolated factor, be it the wisdom of crowds, the competence of the administrators, etc. More explicitly, our formal model, developed below, suggests that two key features of Wikipedia, (i) the ability of administrators to ban contributors, and (ii) the revert function, which enables contributors to restore a previous, better version of an entry, aid in making Wikipedia entries reliable.

We proceed as follows. We start by listing the various factors that make the epistemic reliability of Wikipedia problematic and justify our focus on disinformers and trolls (Section 2). We identify several possible explanations and highlight the difficulty to adjudicate between them non-formally (Section 3). We introduce a computer simulation of a Wikipedia entry, which includes various factors, in particular contributor actions such as history checks and page reverts, as well as administrative measures such as user bans. The simulation also features contributors with a variety of epistemic and non-epistemic aims – namely regular contributors, disinformers and trolls (Section 4). We then gather the results from simulation runs in a number of contexts, gradually adding obstacles to epistemic reliability and identifying possible solutions to them. Overall, these results suggest a pluralist diagnosis – only a combination of factors of various kinds is able to guarantee its reliability (Section 5). We finally discuss this pluralist explanation as well as the robustness and limits of our results (Section 6).

2. Why so reliable?

How reliable is Wikipedia? What is the proportion of true to false claims it contains? Even if replies to this question exhibit some degree of disagreement, all agree that Wikipedia is more reliable than anyone would have thought initially, and in particular that it is surprisingly close to the epistemic performance of traditional encyclopedias. The classic source on Wikipedia reliability is Giles (2005), which compares its error rates with that of the *Encyclopedia Britannica*. On average, there is about one more error in a Wikipedia entry, and the error rate displays more variance. Moreover, "*Wikipedia* contained more entries than *Britannica* with zero errors, but two *Wikipedia* articles were worse than the worst of *Britannica*'s" (Magnus 2009: 75). Even if Wikipedia may be seen as epistemically inferior to Britannica, the point is that the gap between them is surprisingly narrow.

However, the analysis of Wikipedia's reliability cannot be static, because its entries are constantly modified. As Magnus (2009) further emphasises, the speed with which new errors are corrected should also impact our assessment of Wikipedia's epistemic success. It would be damaging that errors, even if all are ultimately corrected, persist for too long. Moreover, the advantages of Wikipedia may not stem from its reliability. According to Fallis (2011), Wikipedia also fares well with respect to Goldman's (1987) criteria of power, speed and fecundity.[1] Still, we will focus on the problem of explaining Wikipedia's reliability only – focusing on the proportion and persistence of errors – as it constitutes its most intriguing feature.

The reliability of Wikipedia is puzzling for a number of reasons. In what follows, we focus on what we consider to be the three major, partly interrelated problems that a satisfying explanation of Wikipedia's epistemic success must solve.

First, it somehow emerges from the individual contributions of a high number of contributors with unknown expertise – anyone can edit a Wikipedia entry, so the average reliability of most contributors may be low. As a consequence, "the true miracle of Wikipedia is that this open system of amateur user contributions and edits doesn't simply collapse into anarchy." (Anderson 2006: 71). What process could aggregate a crowd of individual ignorance into a collective body of information? Call this the *amateur problem*. This is not to say that no expert ever contributes to Wikipedia. However, if such experts are a minority, we should not expect their contributions to be more persistent than any

[1] "We are also concerned with how much knowledge can be acquired from an information source, how fast that knowledge can be acquired, and how many people can acquire it" (Fallis 2011: 305).

other edit, as anyone may alter or delete them. Of course, the importance of the problem depends on the real proportion of non-experts among contributors, which has not been estimated.

A second, related issue stems from the fact that Wikipedia entries are fundamentally volatile bodies of text. By contrast to collective epistemic enterprises such as science, Wikipedia is constantly open to unsupervised revisions. Entries may be instantly modified by any individual contributor. Wikipedia offers no peer review process and no selection of contributors who may revise an entry. As a consequence, given the high number of contributors, most entries are modified on a daily basis and are thus unstable. Moreover, the problem is not only the sheer *quantity* of revisions but their *scale*. The content of an entry may be substantively or totally deleted without difficulty. Indeed, massive deletions have always been frequent in Wikipedia (Viegas et al. 2004). An aggravating factor is the asymmetry of effort needed to contribute positively or negatively: it is easier and faster to suppress some content than to add it (which takes time, if only to write it all). Not only are revisions frequent; we should expect them to be so, given this effort asymmetry. Overall, the quantity and scale of revisions constitutes what we call the *volatility problem*. It partly stems from the amateur problem: many unskilled contributors naturally allow for many errors in revisions and deletions. But regardless of its causes, the volatility problem is fundamentally epistemic: it supposes that content is only informative if it is present during a large enough proportion of time. The fact that informative content tends not to disappear (or reappears often enough) thus begs for an explanation that goes beyond the amateur problem.

A third issue stems from a characteristic of Wikipedia that make it stand out among more typically discussed cases such as epistemic enterprises such as science. As stated in the introduction, studies of science in social epistemology have traditionally considered it as based on individuals that either aim for epistemic success or whose motivations are shown to indirectly favour epistemic success. For instance, Kitcher (1990) compares collective epistemic success (the probability of obtaining a result) in the case of 'pure', truth-seeking scientists and of scientists motivated by reputation or personal gain; Strevens (2003) further stresses the beneficial epistemic consequences of egoistic motives. In Weisberg & Muldoon's (2009) 'epistemic landscape' approach, scientists are attracted by epistemically significant results.[2] Zollman (2007) considers agents who choose theories that appear to be best

[2] Some of these works have been later criticised for their lack of robustness or artificial assumptions. However, here, we are merely concerned with the assumptions of epistemically beneficial agents that are shared by these works as well as their refinements.

supported by the available evidence; and so on. All in all, agents in such formal models or simulations are epistemically beneficial.[3]

Although perfectly fine on their own, the lessons drawn from such approaches cannot be adduced in order to understand Wikipedia. This is because Wikipedia contributors have a variety of motives, some of which are epistemically detrimental. Some contributors are disinformers: they intentionally insert falsities or delete true claims, for instance because of vested interests, personal opposition to a certain view, etc. Disinformers have anti-epistemic aims – they squarely oppose epistemic success. Other contributors have non-epistemic, or a-epistemic aims, that is, their contributions stem from motives whose results are not correlated with epistemic success. For instance, some contribute only to have fun or to waste other people's efforts. Following Frankfurt (2005), such contributors may be called bullshitters, or to the Internet jargon: *trolls*. Disinformers and trolls give rise to the *vandalism problem*, that is, the fact that a significant number of Wikipedia contributors are actively involved in the modification or deletion of true content and/or the insertion of false content.

Put together, these three problems make Wikipedia's reliability intriguing. How could Wikipedia's epistemic success come close to that of traditional encyclopedias, given that its content is highly unstable, most of its contributors neophytes and some of the rest disinformers and trolls? How can Wikipedia be reliable in the face of contributor amateurism, content volatility and epistemic vandalism? We now turn to possible explanations. Note already that the number and variety of problems may make it look unlikely that a unique remedy could solve them all. But this verdict would be premature: we are dealing with a complex system, about which any claim may be difficult to assess in an analytical fashion.

3. Hypotheses

As the use of Wikipedia is widespread, concerns regarding its reliability abound, but so do tentative explanations of this reliability. In what follows, we introduce a number of them. Our aim is twofold. First, we want to highlight the sheer number and variety of types of possibly relevant factors or processes. Second, our list of explanations should make clear that none of them is fully convincing at first glance, and that any assessment of possibly relevant factors is unlikely to be reached

[3] In the non-formal literature, an earlier similar thesis is that of Hull (1989), who argues that the success of science cannot be properly explained if one neglects the scientists' desire for recognition.

analytically or non-formally, which in turn motivates our appeal to a computer simulation.

The first explanation of Wikipedia's reliability is based on the so-called '*wisdom of crowds*' phenomenon (Surowiecki 2004). It is usually invoked when the aggregation of individual outputs is better or more accurate than any of them considered in isolation. One explanation of this phenomenon is that individual errors may 'cancel out' one another, so that even if all individuals are mistaken, the sum or aggregation of their opinions becomes less so. However, it is far from clear that this phenomenon is responsible for the reliability of Wikipedia. The wisdom of the crowds typically happens in a context in which individual assessments are aggregated into a unique collective assessment. By contrast, Wikipedia entries consist in the juxtaposition of individual contributions that concern various aspects of a topic. Part of entries can rarely be seen as averages of multiple individual contributions, and the errors contained in one part of an entry do not compensate those found in another one – all draw reliability lower. Moreover, the wisdom of the crowd is typically thought to happen when individual opinions are independent (Surowiecki 2004). This condition may be met for some topics but not others, especially when the influential sources of information are few. Overall, it is difficult to understand to what extent the wisdom of the crowd phenomenon may justify Wikipedia's reliability.

Another explanation of Wikipedia's reliability relies on empirical properties of the population of contributors, especially on their distribution. For instance, the *elite hypothesis* holds that there exists a small proportion of highly reliable, highly active contributors, which would counterbalance the actions of the vast majority of more passive and less reliable contributors. In other words, the reliability of Wikipedia would be safeguarded by a benevolent, efficient elite. The actual number and distribution of reliable contributors is a purely empirical matter (and it does seem that such an elite exists; see Sanger 2009); but the question of their effect on Wikipedia's overall reliability is not. Would a reliable elite be able on its own to decrease the volatility of entries? To stave off disinformers and trolls, which are thought to be quite active as well? How reliable exactly would an elite need to be in order to effectively counterbalance a mass of neophytes?

Relevant aspects of the population of contributors also include the distribution of the anti-epistemically or non-epistemically inclined ones, that is, of disinformers and trolls. Obviously, the lower the proportion of disinformers and trolls in the population, the more reliable Wikipedia will be. So maybe disinformers and trolls are just too small a minority to substantively disrupt Wikipedia. Again, this explanations hinges on

an empirical fact, but also on an assessment of the extent to which disinformers and trolls make reliability plummet.

While the previous explanations involve characteristics of the population of contributors, other ones focus on technical features of Wikipedia. A Wikipedia entry is not just characterised by its content at a given time; the history of the changes it underwent are public. Although Wikipedia entries are constantly edited and are thus fluid in a sense, the sequences of past edits are publicly available and can always be checked by anyone. Edits are transparent, and so is the list of the past contributors – even if they are anonymous, their IP addresses are available.[4] This allows one to detect whether an entry has been changed multiple times and in similar ways by a given contributor, for instance, or whether it results from the independent contributions of multiple authors. Of course, the extent and nature of the changes are also transparent. So-called 'edit wars' – sequences of opposite changes due to two or a handful of contributors – become easily detectable as well. According to this *transparency hypothesis*, this feature of Wikipedia goes a long way towards explaining its reliability.

Moreover, not only can past edits be known, they can also be reverted. That is, an entry can simply be returned to one of its former states, without any need to rewrite it from scratch. This facilitates the fight against disinformers and trolls, for instance, as it seems to compensate for the effort asymmetry between positive and negative contributions (see section 2). Finally, note that the transparency of entries is further compounded by the existence of public discussion pages, in which the contributors freely discuss entry-related issues – its evolution, the changes that have been or may be brought, etc. One may thus come to know whether given parts of the entry content are consensual or divisive, whether they are considered as being targeted for manipulation by disinformers or trolls, among other things. All these features participate in the transparency of the page and the associated ease with which non-epistemically inclined contributors may be staved off.

Still a different set of explanations of Wikipedia's reliability involves its structure and policy. There exist sets of recommendations regarding the acceptable form of entries. For instance, an entry should be adequately sourced – its affirmations should always be backed up by an identifiable source of information. Other recommendations include the necessity to retain a neutral point of view, the prohibition of original

[4] Moreover, there exist a watchlist tool, which allows users to be notified whenever a page of interest is modified.

research.[5] Such recommendations are part of Wikipedia's policy; accordingly, the *policy hypothesis* holds that they explain a substantial part of Wikipedia's reliability. Of course, they do not prevent contributors from making changes that do not follow them, but provide guidelines that help identify possibly problematic entries. Indeed, sections of entries that do not respect the guidelines are typically flagged for attention, that is, signalled by a tag that may indicate various issues: copy and pasted parts, non-neutral viewpoints, absence of sources, improper references, factual inaccuracy, excessive length, etc.[6]

A final explanation to consider is the *administration hypothesis*. According to population-level hypotheses, the reliability of Wikipedia naturally emerges from the aggregate behaviour and characteristics of its contributors. The administration hypothesis adopts an opposite standpoint by claiming that the reliability of Wikipedia is chiefly explained by some of its high-level features, or by some top-down influences, namely that of its administrators. Wikipedia does not merely result from the combined actions of identical contributors. Some of them are granted an administrator status, which allows them to perform a number of specific actions, such as deleting pages, protecting them from editing, and 'rollbacking' them – reverting them to their previous states much faster than regular contributors. Most importantly, administrators can also ban or unban contributors (identified through their IP address), that is, prevent them from further modifying entries. If contributors could not be banned, disinformers and trolls may keep acting constantly, thus hampering the reliability of Wikipedia by making its volatility skyrocket. The possibility that contributors be banned limits the epistemic pollution due to disinformers or trolls: it motivates them to be subtle or to perform smaller edits than they otherwise would.

We end up with at least five hypotheses pertaining to the reliability of Wikipedia: the wisdom of the crowd hypothesis, the elite hypothesis, the transparency hypothesis, the policy hypothesis and the administration hypothesis. These are not mutually exclusive, as several factors may favour similar results. As all hypotheses share some intuitive appeal, it is difficult to find further grounds for excluding or selecting some of them, or some of their combinations. In what follows, we appeal to a computer simulation of Wikipedia, in order to tease out the respective causal effects of a number of relevant factors. Our results will indicate that the reliability of Wikipedia may stem from a combination of factors without being explainable by any isolated one.

---

[5] The recommendations may be found on:
$https://en.wikipedia.org/wiki/Wikipedia:Verifiability$

[6] See $https://en.wikipedia.org/wiki/Wikipedia:Contributing_{t}o_{W}ikipedia$

4. Simulating Wikipedia

4.1. The model

Many factors may be responsible for the reliability of Wikipedia. In order to have a better grasp of their respective causal influence, we appeal to a computer simulation of Wikipedia. The aim is not to reproduce Wikipedia's functioning exactly (how could one?), but to use a simplified model that includes what we have previously identified as its main features, in order to study how their variations impact epistemic reliability. As the model is unusual and has more moving parts that the ones typically met in traditional social epistemology, it is worth spending time describing it. This also allows us to clarify our modelling assumptions.

The model aims to represent the dynamic evolution of a Wikipedia entry. An entry is represented by a finite list of information units expressing propositions, each of which may be true or false. Information units are also more or less esoteric, that is, their truth value is more or less difficult to detect (a quantitative level of esotericity is associated to each unit).

Entries get modified by a population of users. Users may do three things (at random): to contribute (add an information to the entry), to check information units in the entry or to check its previous version. To each user is attributed a level of activity (the probability that he acts in a given round), as well as a reliability, which represents both the probability that a unit added by the user is true and his ability to detect false information units. The more reliable the user, the more true (and more esoteric) units he may add, the better he is at detecting falsehoods. At each round of the simulation, depending on their activity level, users stay still or choose an action and perform a number of such actions.

Users may be of three different types: honest users, disinformers and trolls. Honest users add information units, the truth of which depends on a user's reliability; but disinformers always add false information. (So honest users may add false information units, but disinformers may not add true ones). Trolls, which are supposed to make random contributions, always add false information units. They can also act more often than other users at each round, which reflects the typically superior activity level of trolls due to the asymmetry of effort described earlier – the fact that it is easier and faster to add any content whatsoever than to contribute either honestly or deceptively.[7]

[7] In particular, one salient characteristic of troll actions is that they are particularly repetitive; see Shachaf & Hara (2010).

When checking, honest contributors check a number of (randomly determined) information units and delete them if they have detected their falsity; similarly for disinformers with units they have identified as true.[8] Trolls delete every information unit they check, regardless of their truth value.

Users may also check a number of previous versions of an entry, which means they check every added or deleted information (as compared with the current version). An honest user will choose the version which is more reliable, that is, whose difference between the numbers of true and of false information units is higher, if there is one, or else keep the current version; similarly with disinformers who choose the less reliable version. Trolls revert to a random version, within a higher pool of past versions than honest users.

Finally, our model includes an administration. When honest users check previous versions, they may report disruptive edits to an administrator. Disruptive edits involve additions of falsities, deletions of truths, reversions to less reliable versions of an entry, or mass modifications (which only trolls can perform). The user responsible for the edits is then banned if these were actually disruptive (and not just considered so by the user).

This leaves us with many factors to consider. Different distributions of user reliability or activity are possible. The proportion of action types (contributions, checks, reverts) has to be chosen, as well as the number of actions for each user type (honest, disinformer, troll) and the distribution of user types.

4.2. Results

We now describe the results of a number of simulations run for diverse populations of contributors; we illustrate and interpret the main ones, while only mentioning others in passing. Fig. 1 illustrates what happens with a homogenous population of honest contributors: the quantity of true information grows steadily, while that of false information remains low; the page converges towards a high reliability. Moreover, this holds even if the users' reliability distribution has a very low average – average reliability only impacts the speed of convergence.[9] Similarly, increasing

[8] Note that in our model, this process involves noise. More precisely, the model compares the reliability of the checking user, modified by some noise, to the reliability of the author of the checked information, also modified by some noise. When the former is greater than the latter, a user will know whether an information is true or false.

[9] For instance, with a gaussian distribution of 0.1 average, the entry's reliability still converges to 1, although more slowly, because of two factors. First, the rate of increase of true information is about a third of what it was for the 0.5 average

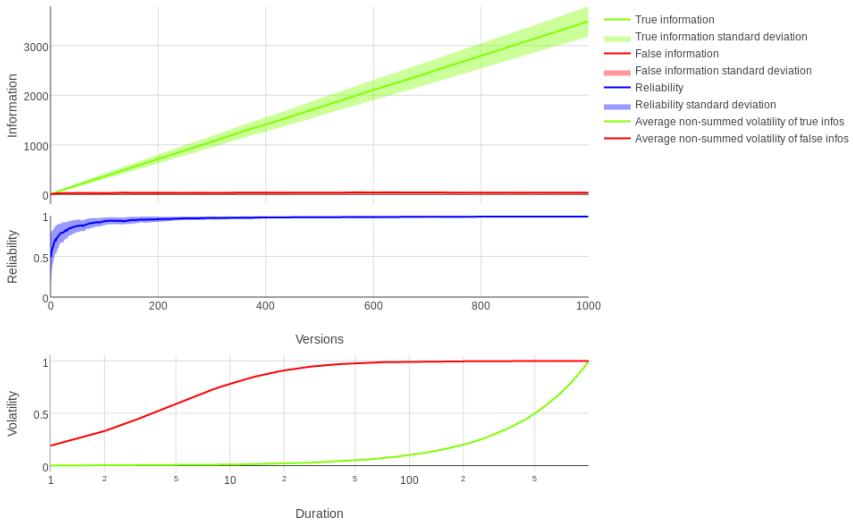

Figure 1. Homogenous population of honest users: overall quantity of true and false information, average reliability of information, average volatility of true and false information.

the proportion of check/delete actions as compared to regular contributions, increases the speed of convergence (but reduces the total quantity of information contributed).

Note that this should not be interpreted as a solution to the amateur problem. This is because true information cannot be deleted, and so necessarily accumulates as all users are honest. Still, the result would hold even if honest users were allowed to delete true information but only rarely did so, that is, if they were themselves reliable enough. In other words, the entry's reliability stems straightforwardly from the users' characteristics.

Why not model more realistic users? This is because we aim to focus on disinformers and trolls. As we will see, they are powerful enough to counter the positive effects even of such ideal honest users, which in turn suggests that they constitute the main issue for the reliability of Wikipedia.

The introduction of disinformers shakes things up (Fig. 2 – results obtained with a 50% proportion of disinformers). The quantity of false information now steadily increases, although its variance does as well.

condition. Second, there is a constant, noisy but non negligible amount of false information.

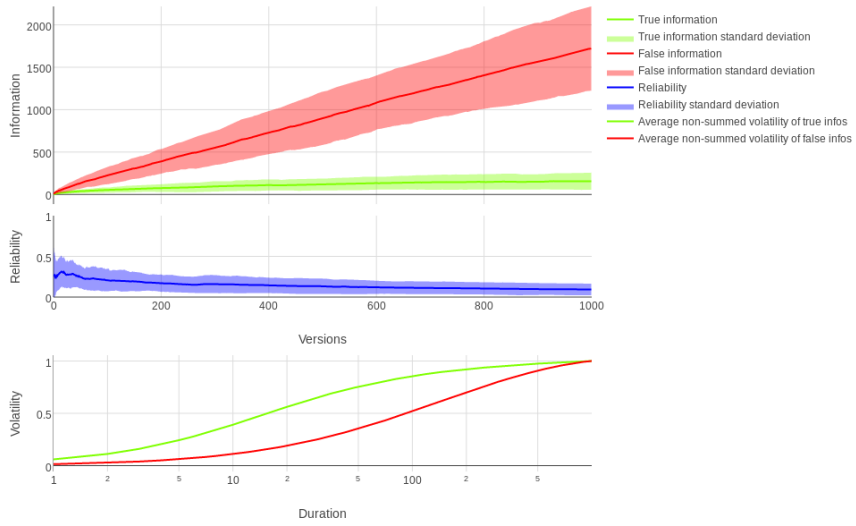

Figure 2. Heterogenous population (50% honest users, 50% disinformers).

True information accumulates much more slowly, because it is more volatile. Accordingly, the simulation converge towards a low reliability.

Trolls are even more disruptive than disinformers (Fig. 3), because they prevent any convergence and create an extremely high variance of information. False information dominates true information, although without increasing, and reliability is low on average. But the variance of false information and so of the entry's reliability is high. Because trolls do not care about the truth value of the information they modify, true and false information are almost equally volatile (the difference is due to the actions of honest users).

So far, the results suggest that the model adequately represents the intuitive effects of disinformers and of trolls. The average reliability is lower with disinformers, because false information accumulates. Trolls prevent any such accumulation, but affect even very esoteric pieces of information. In addition, for a given proportion – even a low one – trolls are more harmful than disinformers overall. A 25% proportion of trolls still prevents convergence towards a high reliability; such convergence, although possible for a 5% proportion, still allows for occasional massive deletions (more on this below).

Can users fight against disinformers and trolls ? Increasing the proportion of check/delete actions for honest users is mostly ineffective: it only slows down the negative effects of the disinformers, without

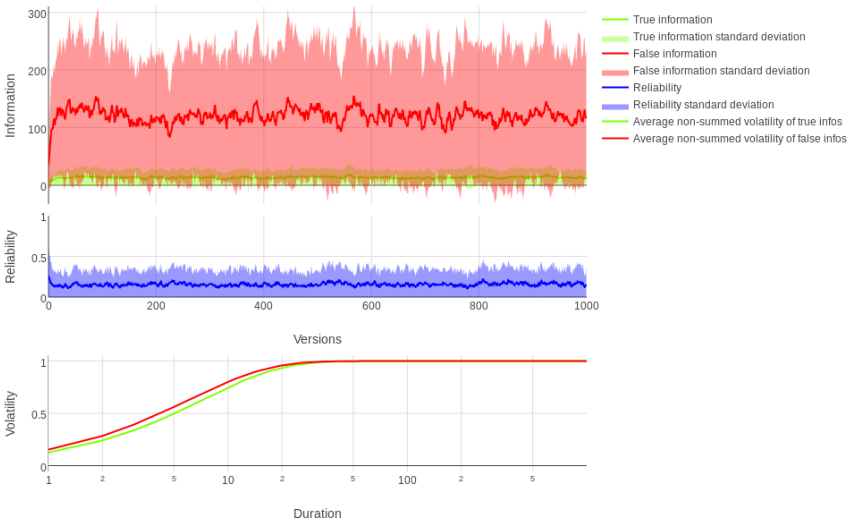

Figure 3. Heterogenous population (50% honest users, 50% trolls).

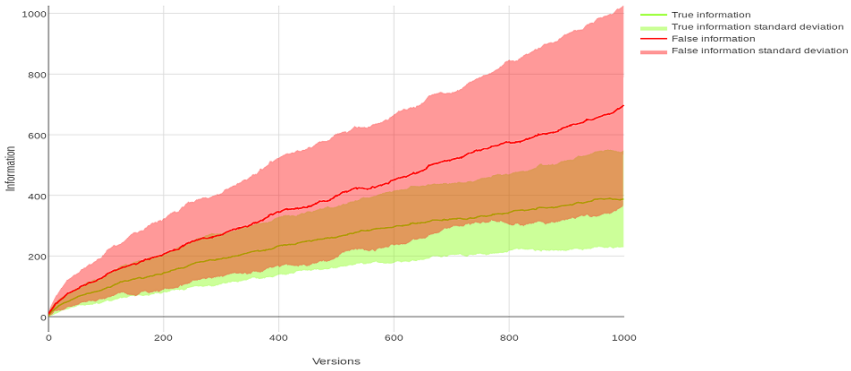

Figure 4. Heteregenous population (50% disinformers, 30% normal honest users, 20% elite honest users).

affecting trolls. Put differently: the wisdom of the crowd does not withstand the disruptive effect of negative epistemic motivations. What about the presence of a number of elite contributors, with significantly higher reliability than honest users? This is mostly ineffective against trolls, and against disinformers only leads to a slow increase (and high

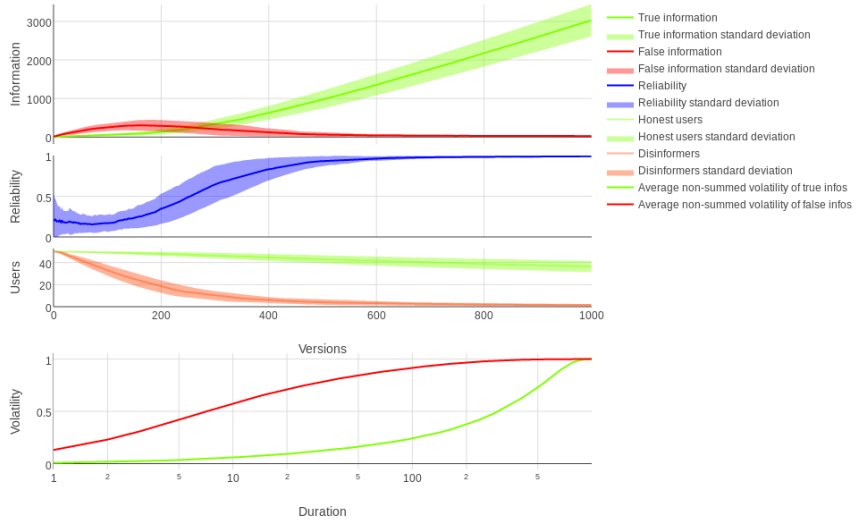

Figure 5. Heterogenous population (50% honest users, 50% disinformers) with administration.

variance) of true information, which still pales when compared to that of false information (see Fig. 4).

Let us now introduce administration, which is able to ban some authors of disruptive edits if they are flagged by honest contributors. Does it suffice to stave off disinformers? It does (Fig. 5). Among other things, administration targets those who add falsities and/or delete truths. Honest users may add falsities, and are sometimes banned; but disinformers perform such actions more often and so are banned faster. This allows their proportion in the population to decrease so that convergence towards reliability is ultimately restored.

However, administration only partly harms trolls. In the long run, it successfully reduces their proportion; however, even small proportions of trolls are disruptive, as Fig. 6 makes clear. A 5% proportion of trolls is sufficient to regularly generate mass deletions of an entry, which thus has to be rebuilt from scratch. In other words, small numbers of trolls are damaging enough to guarantee a serious volatility problem. This is understandable because trolls are typically much more active than honest users – their actions are less costly and so can be performed more often for an equal level of effort.

Trolls turn out to be resistant even to large communities of honest users, even if policed by an administration. But what if the action

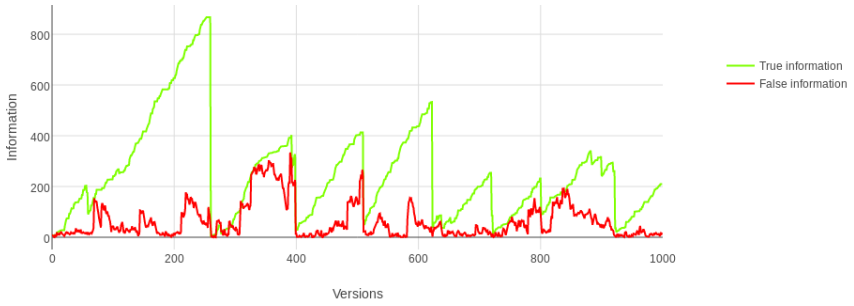

Figure 6. Information volatility for 5% trolls.

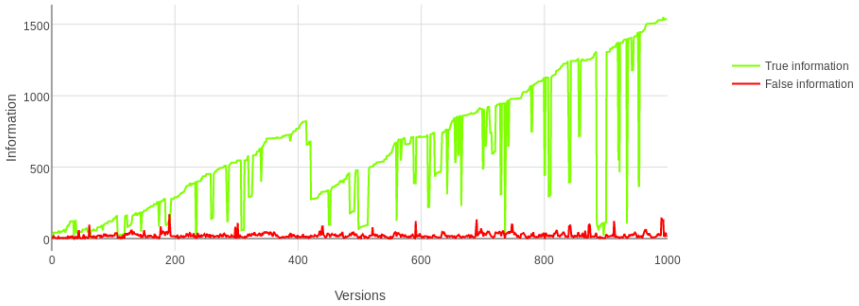

Figure 7. Information volatility for 5% trolls and with the possibility of reversion.

palette of the honest users is enriched so as to reduce the asymmetry of effort between them and the trolls? This is precisely what the revert action permits – recall it allows one to instantly replace an entry by one of its older versions.

Unfortunately, the revert option alone is not up to the task of countering agents with negative or non-epistemic motivations. First, it does not prevent the convergence towards unreliability produced by disinformers; its main effect is to decrease the volatility of information. Second, it is only effective against a small enough proportion of trolls, in which case the entry becomes stable, because mass deletions are typically reverted, and converges towards a high enough reliability (see Fig. 7). However, if numerous enough, trolls will render an entry highly unstable even if the honest users have a revert option available. This is simply because the high activity of the trolls swamps the corrective effects of

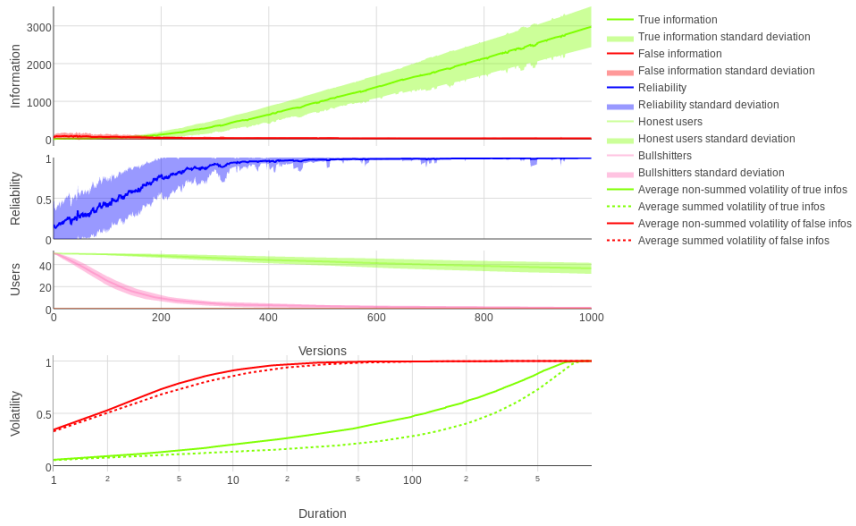

Figure 8. 50% trolls, administration and reversion.

the honest users. Overall, the efficiency of the revert option against trolls depends on their proportion, which is an empirical factor.[10]

Where are we now? An administration counters the disinformers and reduces the number of trolls; the possibility of reversion counters small proportions of trolls. One obvious solution is thus to combine them (Fig. 8), which indeed restores the convergence of the entry towards high reliability. In the beginning, true information is rare and reliability low with a high variance. Then, the administration starts banning disinformers and trolls, thus reducing their numbers; it also bans some honest users, although at a lower rate. Ultimately, only a small proportion of trolls remains, but the possibility to revert the entry ensure its resilience against their regular mass deletions.

[10] Moreover, the threshold proportion of trolls from which reversion becomes inefficient is typically around 10%; but this value varies depending on parameters such as the size of modifications performed by trolls.

5. Discussion

5.1. Empirical adequacy

Even if several explanations of the reliability of Wikipedia have been proposed, as seen in section 3, the successful epistemic consequence of aggregating diverse, possibly unreliable users, is often cited as one of its main causes. By contrast, our model emphasises the importance of structural and technical characteristics, that is, of factors both beyond and below the individual level.

Perhaps surprisingly, Wikipedia's administration plays a major role by banning troublesome contributors and maintaining their proportion under a manageable threshold. It is difficult to estimate the proportion of disinformers and trolls who contribute to Wikipedia; however, it is inferior to 20% – the estimated proportion of "vandals", that is, of contributors whose contributions are all reverted by others (which includes honest but unreliable ones – see Tsvetkova et al. 2017).

Similarly, it would be easy to underestimate the effect of a simple technical option such as reversion. Indeed, it has rarely been identified as a key factor for the success of Wikipedia, although it is commonly performed.[11] Our model suggests that it may play the key role of rendering innocuous the trolls not yet banned by the administration, by reducing the asymmetry of effort between them and the honest users.

Our simulation results thus points towards the possibility of a pluralist explanation for the reliability of Wikipedia, in which factors of different kinds play complementary roles: the behaviour of honest users matters, but so do the specific technical options available to them and the administration structure to which they can appeal.

Of course, we only claim this particular scenario to be a possible explanation for the reliability of Wikipedia, not a necessary or even a plausible one. For it is difficult to argue that simulation runs based on a particular model have any confirmatory power regarding a particular hypothesis. Still, we think that the possibility of such a multifaceted scenario, in which the respective roles of various factors are clear, is interesting in itself.

How does the model fit the real Wikipedia? The number of model parameters which one can adjust makes any reply difficult. Moreover, many of these parameters correspond to user characteristics that are hardly observable (such as reliability, but also the respective proportions of checking and of contributing, for instance). However, such an assessment is not out of reach.

[11] According to Viegas et al. 2004, about half of mass deletions are reverted within three minutes.

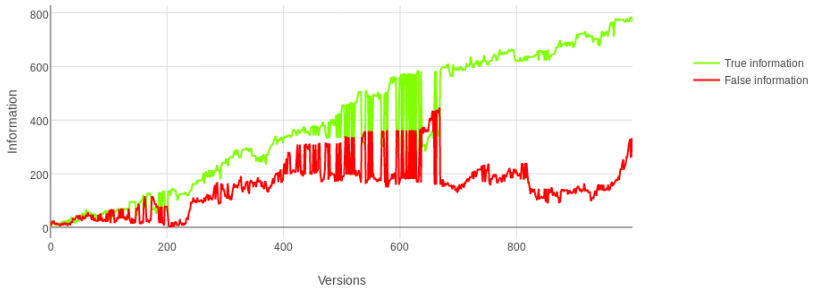

Figure 9. One example of an edit war, here starting shortly after version #400 (60% honest users, 40% disinformers, no administration).

First, recall that the final results, represented in Fig. 8, are only obtained when the proportion of trolls is (roughly) below 10%. Our explanation of the reliability of a Wikipedia entry thus wouldn't fit cases in which a higher proportion of trolls are active. If, as suggested above, the actual proportion of disinformers and trolls and Wikipedia is below 20%, then it is plausible that there are less than 10% of trolls – because it is plausible that a majority of dishonest users are disinformers, trying to propagate false information, rather than trolls, who practice epistemic vandalism for the sake of it.

Second, even if it is difficult to extract predictions from the model, it is notable that it allows for the emergence of a commonly observed Wikipedia phenomenon: edit wars. Edit wars correspond to succession of massive changes and/or reverts in opposite directions – in other words, episodes in which a sizeable amount of information is deleted, then restored, then deleted again, and so on.[12] We take it as a virtue of our model that it shows edit wars taking place from time to time; Fig. 9 provides an illustration.

That an existing Wikipedia phenomenon can unexpectedly appear in our model is already a positive sign. In addition, the empirical adequacy of the conditions under which edit wars may emerge in our model can thus be explored. For instance, in the simulation edit wars typically happen when the proportion of disinformers is close to that of honest users, and when trolls are nearly absent. Edit wars stem from disinformers. This is because troll-based reverts are content-independent and so shouldn't be expected to repeatedly target the same information. Moreover, the frequency and length of edit wars typically increases as users activity involves a higher proportion of version checks. All

[12] See $https://en.wikipedia.org/wiki/Wikipedia:Edit_warring$

these features are observable in principle and so may help provide an empirically assessment of the model.

5.2. Robustness

Our results are robust with respect to a number of factors. Changes in the reliability distribution (whether uniform or normal with various means) of users only affects the speed of convergence; so do modifications of the number of actions allowed for all users. Increasing the number of actions of trolls affects their disruptive power but not the qualitative results.[13] One change of consequence concerns the proportion of original contributions and of checks for honest agents: more contributions entail more information but a decreased reliability; more checks lead to less information but a higher reliability, which is understandable.

Of particular interest is robustness with respect to the activity profile of various users. There is evidence that user activity is very unequally distributed: a small number of editors make a disproportionate number of modifications,[14] which may be thought as a partial explanation of the reliability of Wikipedia. However, the model results turn out to be robust with respect to activity distribution. For instance, if user activity follows a Pareto distribution (that is, if it is disproportionately concentrated in a small portion of the population of users), we obtain results that are indistinguishable from that of Fig. 8.

Note however that this does not allow us to draw general conclusions, for at least two reasons. First, user activity may be correlated with high reliability: if the most reliable users happen to contribute disproportionately more than unreliable ones, then there may be consequences regarding general reliability (which could possibly support the elite hypothesis). Second, very active users typically contribute to many entries, disproportionate activity distribution may help explain the reliability of Wikipedia as a whole without playing a huge role a the level of an isolated entry, which is what our model represents.[15]

That being said, at least one aspect of the simulation is not particularly robust: the volatility measure. This is because the volatility of a piece of information, that is, the total duration of its presence online, crucially depends on specific parameters. For instance, the proportion of honest users as well as the proportion of checks they perform ty-

[13] Note that the results would not necessarily be robust to a *decrease* in troll activity, because this would both make trolls less disruptive and harder to detect (which hampers the efficiency of the administration / revert solution).

[14] See
$https://en.wikipedia.org/wiki/Wikipedia:List_of_Wikipedians_by_number_of_edits$

[15] We thank an anonymous referee for drawing our attention to the distribution of user activity.

pically impact volatility measures. However, note that this does not prevent one from comparing various volatility values for different scenarios after such parameters have been fixed, which is precisely what we have done. Moreover, this does not qualitatively affect the results.

5.3. Modelling choices

There is at least one serious objection that may be levelled against our model, which concerns the idealisation of the behaviour of honest users. To recall, in our model honest users add information which may be true or false (depending on their reliability) and delete information which they have detected as false. In particular they do not delete information whose truth value is unknown. Let us label this characteristic as *strong epistemic honesty*. One may argue that it is excessive demand and that honest users should only display weak epistemic honesty, by being allowed to err also when they delete information. Surely, even well-meaning individuals sometimes mistakenly delete true information.

However, modelling this would introduce more arbitrary parameters in the model. One obvious possibility would be to introduce an uncertainty threshold: when a user has not detected the truth value of an information but came close enough, then he would perceive a truth value (to be determined either randomly or on the base of the user's reliability) and delete it if perceived as false. This would necessitate a choice for the value of such a threshold, in a model that already has many moving parts. We thought such arbitrariness too costly, given the fact that there are independent reasons why strong epistemic honesty would suffice.[16]

For apart from this technical point, there are at least three possible rejoinders to the objection that honest users should be modelled as displaying weak rather than strong honesty. First, if the empirical frequency of mistaken deletions is low, then our model may still approximate the dynamics of Wikipedia reasonably well. As it happens, there are reasons to consider this frequency to be low, which involve the Wikipedia's verifiability policy. While the addition of non-sourced information is a frowned upon but tolerated practice, the deletion of sourced information is detectable, revertable and typically punished. This implies that honest contributors should be more wary and so less likely to delete true information than to add false information – they are pushed towards strong epistemic honesty. Note that this line of argument thus brings into the mix an explanatory factor of yet a different kind, that is, a policy-level one.

[16] We than an anonymous referee for urging us to mention such motivations.

The second rejoinder is that a homogenous population of weakly epistemic honest users may be considered as equivalent to a heterogenous population of strongly epistemic honest users and of disinformers (which may delete true information). While the respective behaviours of such agents differ in a given run, statistically they amount to the same effects.

The third rejoinder is different. Considering only strongly epistemic honest users both highlights the problem posed by disinformers and trolls and strengthens the relevance of our pluralist explanation. It is testimony to the disruptive power of disinformers and trolls that they can have dire epistemic consequences even within communities that contain numerous strongly epistemic honest users. Moreover, if only a mix of various factors justifies the reliability of Wikipedia in an idealised community, then individual factors are even less likely to guarantee similar results in isolation. If it takes a combination of factors to push idealised users towards collective reliability, then it shouldn't take any less to push less ideal ones. Considering strongly epistemic honest users thus buttresses the need for a pluralist explanation of the reliability of Wikipedia.

5.4. Limits of the model

Our simulation model is idealised in a number of other ways as well; let us mention four of them. First, it only represents the evolution of one entry, whereas Wikipedia contains millions of them.[17]. Multiple entries allow for multiple reliability values for each user, depending on the topic considered. Multiplying entries may thus alter our core results at least if there are interaction effects between different entries. For instance, a user suspected of disinformation or trolling regarding one entry may be banned even if its contributions to other entries were legitimate; so if users can be honest regarding some topics and dishonest regarding others, then relatively to one entry, honest users would be more affected by the administration than in our model. However, it is not clear that the results would not be only quantitatively but also qualitatively different.[18]

A second idealization concerns the evolving composition of the user population. In our model, the proportion of user types are fixed in each simulation. However, it is possible that these vary as time passes.

[17] Currently more than 5 millions – see *https* : *//en.wikipedia.org/wiki/Wikipedia* : $Size_of_Wikipedia$

[18] One particular difficulty for modelling multiple entries is to find a common time scale. In our graphs, changes are measured as functions of the number of changes of entry version. However, such changes may occur at different paces for different entries, which complicates their combination.

For instance, trolls may drive off honest contributors, tired of endlessly restoring their contributed content. This echoes Sanger's (2009) claim that the reliability of Wikipedia should decrease in the long run because good contributors are discouraged by aggressive behaviours, themselves magnified by user anonymity. In general, it would be interesting to model and explore the negative influence of disinformers and trolls on the motivation of honest users.

Third, the technical options offered by Wikipedia have only been touched upon. As powerful as the revert option is, it is but one of the tools that users have at their disposal. Discussion forums, for instance, allow them to debate the status and evolution of a page before deciding which course of action should be taken. It is difficult to see how this should be modelled though. Another action available to the users is the flagging of various parts of an entry, in order to signal some of its issues (e.g. lack of referenced sources, excessive length, etc.).

Finally, the administration level of Wikipedia too is more complex and layered than in our model. Wikipedia does not just stem from user contributions and administrators with a banning power. Administrators may also protect, restore or delete pages and typically form judgements from the discussions that takes place in forums. They may also be sources of abuse themselves; their actions may be reverted and their status removed.[19] Still, such measures are rare and most admins do regularly ban problematic users.

We conclude this section by discussing the explanatory power of our model. Does it allow us to draw conclusions regarding collective epistemic systems other than Wikipedia? In particular, does it provide reasons to think that explanations of the reliability of such system should be pluralist as well?

Intuitively, the explanatory factors we highlight should be exportable to the extent that the model characteristics are shared by other epistemic systems. On the one hand, many online collective epistemic systems (forums, databases, etc.) are run or supervised by administrators. On the other hand, the Wikipedia revert function, which in our model is crucial for high reliability, is typically not instantiated similarly in other systems. Online databases always include backup versions to which they can be reverted in case of a problem (e.g. hacking). This, however, depends on a decision taken at the administrative level, not by the database contributors themselves. As a consequence, it is not clear whether the epistemic advantages that the revert option brings to Wikipedia, in which it is distributed, would also be present in al-

[19] For a fuller description of Wikipedia's administrators, see *https* : *//en.wikipedia.org/wiki/Wikipedia* : *Administrators*

ternative systems in which the revert decision is centralised. Overall, our model results may be generalised to the extent that a distributed revert option is available in the system of interest.

An additional conclusion is that reliability explanations for collective epistemic systems may crucially depend on local properties of such systems. While the distributed revert option is key to the reliability of Wikipedia, that of alternative systems may depend on different technical features, the efficiency of which itself hinges on the kind of anti-epistemic agents who tend to modify them. In other words, there may not be any general explanation of the reliability of collective epistemic systems, but only local ones.

6. Conclusion

The surprisingly high reliability of Wikipedia has often been seen as a beneficial effect of the aggregation of diverse contributors, or as an instance of the wisdom of crowds phenomenon. Moreover, all potential explanatory factors of this reliability have only been defended non formally. Our aim was to assess such explanations in order to further understand the reliability of a collective knowledge-forming process that is seldom studied in social epistemology: namely, one in which individual agents have non and/or anti-epistemic aims.

We have provided a first computer simulation of a Wikipedia entry, which has allowed us to assess the respective role of factors of different kinds. We identify the main threat for Wikipedia as the presence of negative or non-epistemically motivated contributors, namely disinformers and trolls, which honest contributors cannot be expected to counterbalance.

In our model, the reliability of a Wikipedia entry turns out to stem from the combination of at least two crucial factors: the administration and the revert option. By banning damaging users, the administration is able to stave off disinformers and to reduce the number of trolls; the disruptive effect of the remaining small proportion of trolls is offset by the use of the revert option by honest users. The reliability of Wikipedia is thus due to factors of at least three kinds – individual (honest users), structural (Wikipedia's administration) and technical (the revert option) – none of which would be sufficient in isolation.

Our scenario provides but a possible, and hopefully plausible explanation for the reliability of Wikipedia. However, it suggests that the correct explanation – whatever it is – may be a pluralist one, that is, it may rely on the interaction of various factors rather than amount to a single reliability-inducing process. Wikipedia is a complex system; there

is no reason to expect a single factor to be responsible for its success. However, it is not so complex as to be impervious to modelling. We hope this work will pave the way for more sophisticated models that will shed further light on the recent and strange phenomena of online collaborative knowledge.

Acknowledgments

We thank Anouk Barberousse for her comments on an early version of this work.

References

Anderson, C. *The Long Tail*. Hyperion, New York, 2006.

Fallis, D. Wikipistemology. In A.I. Goldman and D. Whitcomb, editors, *Social Epistemology: Essential Readings*, 2011. Oxford University Press.

Frankfurt, H. *On Bullshit*. Princeton University Press, 2005.

Giles, J. Internet Encyclopaedias Go Head to Head. *Nature*, 438:900–1, 2005.

Goldman, A. Foundations of Social Epistemics. *Synthese*, 73:109–144, 1987.

Hara, N. and Shachaf, P. and Hew, K. F. Cross-Cultural Analysis of the Wikipedia Community. *Journal of the American Society for Information Science and Technology*, 61(10):2097–2108, 2010.

Hull, D. *Science as a Process*. Chicago: University of Chicago Press, 1988.

Kitcher, P. The Division of Cognitive Labor. *The Journal of Philosophy*, 87(1):5–22, 1990.

Magnus, P. D. On Trusting Wikipedia. *Episteme*, 6(1):74–90, 2006.

Sanger, L. M. The Fate of Expertise after Wikipedia. *Episteme*, 6(1):52–73, 2006.

Shachaf, P. and Hara, N. Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, 36:357–370, 2010.

Strevens, M. The Role of the Priority Rule in Science. *The Journal of Philosophy*, 100(2):50–79, 2003.

Surowiecki, J. *The Wisdom of Crowds*. Anchor Books, 2004.

Tsvetkova, M. and García-Gavilanes, R. and Floridi, L. and Yasseri, T. Even good bots fight: The case of Wikipedia. *PLoS ONE*, 12(2): e0171774, 2017. doi:10.1371/journal.pone.0171774

Viegas, F. and Wattenberg, M. and Dave, K. Studying cooperation and conflict between authors with history flow visualizations. *Proceedings of the Computer-Human Interaction*, 6(1):575–582, 2004.

Weisberg, M. and Muldoon, R. Epistemic Landscapes and the Division of Cognitive Labor. *Philosophy of Science*, 76(2):225–252, 2009.

Wray, K. B. The Epistemic Cultures of Science and Wikipedia: A Comparison. *Episteme*, 6(1):38–51, 2006.

Zollman, K. J. S. The Communication Structure of Epistemic Communities. *Philosophy of Science*, 74:574–587, 2007.