Rosenbaum's Magical Entity:
How to Reduce Misinformation on Social Media

By Avi Tuschman*, August 2020

I. AN INCREASINGLY SEVERE AND URGENT CHALLENGE

Our country passed a grim milestone on May 27, 2020, when the first 100,000 lives had officially been lost to the COVID-19 pandemic. On this same day, Twitter made a fateful decision: for the first time, the social media platform placed a fact-check warning on one of President Trump's tweets. This small act set the stage for a much more direct and high-stakes conflict between prominent social media platforms, government, and public opinion. On the very next day, May 28, the president signed an executive order challenging Section 230 of the Communications Decency Act, which immunizes platforms like YouTube, Facebook, and Twitter from liability for third-party content.

This executive order came just five months ahead of the 2020 presidential election, and after years of Senate and House Intelligence Committee investigations that have revealed significant foreign state-sponsored interference in the 2016 US election, orchestrated primarily by manipulating social media platforms. Russia-sponsored disinformation on YouTube has continued to be a significant problem well beyond electoral years, generating billions of views, posing an ongoing threat to democracy.

To make matters worse, there is now also a COVID-19 infodemic whose primary vector is social media platforms. The head of the WHO, Tedros Adhanom, has warned that "fake news spreads faster and more easily than this virus, and is just as dangerous." These torrents of misinformation have directly caused real-world impacts, including deaths due to non-compliance with medical advice, as well as from poisoning and interethnic communal violence. Such events recall the words of the French Enlightenment writer Voltaire: "Those who can make you believe absurdities, can make you commit atrocities."

Perhaps most concerning of all, only 49 percent of Americans intend on immunizing themselves against COVID-19 if and when a vaccine becomes available, and 20 percent

---

* *Avi Tuschman is a Stanford StartX entrepreneur and a pioneer in developing and commercializing privacy-safe Psychometric AI. Tuschman is an expert on the science of heritable psychometric traits. His book and research on human political orientation have been covered in peer-reviewed and mainstream media from 25 countries. Previous to his career in tech, he advised heads of state as well as multilateral development banks in the Western Hemisphere. Tuschman completed his undergraduate and doctoral degrees in evolutionary anthropology at Stanford University.*

say they will refuse to get vaccinated. In order to achieve herd immunity, at least 70 percent of the population will likely need to be immune. Innovating an effective treatment for the infodemic is therefore nearly as urgent as developing an actual vaccine.

II. THREE TYPES OF SOLUTIONS

As the losses and risks caused by misinformation on social media intensify in frequency and severity, solutions to this challenge are increasingly discussed and implemented. Although there are many ways to mitigate this complex problem, it is helpful to divide potential solutions into three categories.

The first category entails government regulation imposed on social media platforms and their content. In the best case, legislation that regulates social media will be too slow to effectively update and will not easily scale. In worse cases, some governments may seek to fine or extort social media platforms, while using them to spread disinformation and to propagate a post-truth era of confusion ripe for anti-democratic, authoritarian power consolidation.

The considerable public health and economic stress brought about by the pandemic are accelerating the democratic recession, as pointed out by Larry Diamond of Stanford's Freeman Spogli Institute. Francis Fukuyama has also analyzed the rise of populist leaders in many countries, a phenomenon that had been in motion well before COVID-19 but which is now growing stronger. It is clear that democratic institutions are increasingly imperiled by these trends; therefore, protecting the independence of speech moderation on social media platforms from government regulation, legal intimidation, and manipulation is more important than ever.

The second type of approach is for social media platforms themselves to take responsibility for governing the limitations of free speech. With regard to arbitrating factuality, this idea has been rejected by Facebook CEO Mark Zuckerberg, especially within the context of political speech. At the same time, Facebook's new content-moderation Oversight Board, which has widely been referred to as the Facebook "Supreme Court," most prominently embodies the second option more broadly. This entity comprises 20 high-profile academics, jurists, and journalists, among them individuals with a history of political activism. Oversight Board member Alan Rusbridger has described his counterparts as an "interesting bunch of lawyers, human rights activists, academics, journalists, troublemakers. If you wanted a quiet life, I don't think you would have chosen this board."

Twitter CEO Jack Dorsey's stand against certain of President Trump's tweets also fits within this second category. The warning label applied to the President's tweet about violence in Minneapolis makes it very clear that the social media platform itself is regulating free speech: "Twitter has determined that it may be in the public's interest for [Trump's] Tweet to remain accessible."

The perception that a small group of unelected elites may exercise power over free speech on the world's most powerful information platforms – and seemingly with different standards applied to different users at different times – will of course continue to be a lightning rod for the media and public opinion, regardless of the coherence of their policies and the consistency of their enforcement. Rusbridger rightly points out that the Facebook Oversight Board will "be criticised whatever we do," and that it "can't possibly deal with the millions of issues that are contested on Facebook." Movements in only this second direction are likely to accelerate government regulation, including a rollback of Section 230.

If a small group of affiliated judges, a CEO, or specialized executives make editorial policy decisions, such decisions may not be sufficiently proactive; if they are instituted as a Supreme Court of last appeal, then conflicts will already have significant visibility by the time of a ruling. Moreover, enforcing their decisions is unlikely to occur quickly enough to effectively mitigate new misinformation, especially at the considerable scale required. Even if an army of content-moderator contractors attempts to keep up with misinformation, all such actions comprise editorial decisions, and thus the social media platform takes on political liability (and potentially future legal liability) for content.

Attempts by social media platforms to regulate misinformation are indeed vitally important and necessary. However, if these efforts continue to evolve as predominantly internal processes, then they will likely be insufficient and will almost certainly continue to escalate the conflict between several large tech companies and the older powers of the state and public opinion.

III. THE THIRD ENTITY

"This conversation always makes my head explode, because we think that the platforms probably can't regulate themselves, but at the same time we don't really want the government to regulate them because arguably we think the government in some countries, theoretically our own, but certainly others without a doubt, will use that power for the opposite of good – that the web will become less free... So it's like we want this magical entity that isn't the government that isn't Facebook or YouTube or Twitter... to come in and set some boundaries. And I don't know what that entity is."

– Steven Rosenbaum, NYC Media Lab, at SXSW 2020 panel on "Revisiting Section 230"

Steve Rosenbaum, who pioneered the use of User Generated Content, rightly suggests that the solution to the misinformation challenge requires a third entity. Tech companies themselves are best positioned to innovate in this direction, to proactively open up a new space to integrate this third entity.

An optimal solution should significantly lessen deleterious content on social media, while simultaneously reducing a platform's political liability for content. It should deflect the heat of incessant media criticism, and it should deflate pressure to eliminate Section 230 in the US or to enact unfavorable new regulation in other countries.

From a technical perspective, a successful solution must be fast, scalable, and proven to accurately and reliably classify truth from falsehood. The ecosystem should retain both centralized and closed parts, but also enable some limited new decentralized and open parts.

Perhaps most importantly, any feasible solution **must ensure that a social media platform's business model, core algorithms, and as much existing content as required remain absolutely untouched and unchanged**. Such a solution is not only possible – its key components are already well developed, proven, and accessible.

A. THE WORLD'S INFORMATION

Organizing the world's information is a formidable and enormously consequential task. The companies that have successfully undertaken this feat have achieved true greatness. In doing so, they've also attained considerable power, which is now being increasingly challenged by traditional political authorities, both democratic and authoritarian.

Google, YouTube, and Facebook are the world's three top websites outside of China. Their status is not an accident: they organize important sets of the world's information exceptionally well according to popularity-driven algorithms, from the original PageRank to the lookalike audiences that maximize engagement and other digital advertising KPIs.

Popularity-driven Web 2.0 platforms, which now process petabytes of data with increasingly powerful AI, have created many wonderful benefits: connecting people across time and space, spreading news and information, reducing in some ways the power of tyrannical governments, lowering the barrier to entry for talented new content providers, and providing the most powerful advertising technologies ever invented to even the smallest and newest businesses. Efficient advertising is vital for everyone because there are only so many ways in which the economic pie can grow (especially in a future with dramatically shrinking populations).

As we've seen, however, popularity has an uncomfortable and awkward relationship with factuality. Systems that overly optimize for popularity, in the absence of safeguards, have problems with the reliability of information. A recent study published in *BMJ Global Health*, for instance, analyzed 69 of the most viewed English-language coronavirus videos on YouTube on March 21, 2020. More than a quarter of them contained misleading or inaccurate information, and these ones had already received more than 62 million views.

Not all Web 2.0 platforms, however, have evolved to rank content by popularity. The world's fifth website (outside of China) instead organizes the world's information in a very different way: a distributed group of non-employee volunteers use a complex technology to organize the world's information according to reliably documented facts.

This website is Wikipedia. It ranks higher than Amazon, Netflix, and Instagram. By summarizing the world's information in reference to fact-checked sources, Wikipedia has truly achieved one of humanity's greatest accomplishments. It has arguably done far less well, however, at raising awareness among a broad audience as to *how* and *how well* the website functions. Therefore, it is vital to dispel a number of misconceptions, which are common among even the most educated and sophisticated company.

B. SHEDDING LIGHT ON MISCONCEPTIONS

There is a saying, popularized by Harvard Law and CS professor Jonathan Zittrain, that "Wikipedia works really well in practice, just not in theory." What this witty observation means is that Wikipedia only seems not to work well in the folk theory of the popular imagination.

Part of the problem may be that Wikipedia bills itself as "the free encyclopedia that anyone can edit." While this tagline is true, it is woefully incomplete. The phrase does a much better job of conveying the ethos of the Internet movement that inspired it than communicating the treasure of knowledge and the patrimony of humanity that Wikipedia has become. The word "free" belies its immense value, and the word "anyone" belies the accuracy of the end result. As Jake Orlowitz, the founder of the Wikipedia Library, explains, few people

> "…understand that getting 'published' on Wikipedia is like stepping into a gauntlet of both algorithmic and human filtering. An individual edit must pass through targeted text-rejection filters to even make it on the page. Then neural network machine learning bots seek out nuanced patterns of vandalism. After that, thousands of human 'recent change' patrollers look at every suspicious new edit, like a game of whack-a-mole. Over the next few hours and days experienced editors are notified of updates to any article on their 'watchlist', a feed of changes to articles in their specific areas of interest and expertise. At last, words are left for the eyes of millions of readers, many more of whom fix an error than add one."

Thus, Wikipedia has enabled "a commons of public fact-checking," enforcing a Neutral Point of View and attribution to reliable sources that have a reputation for both fact checking and accuracy. Consequently, Wikipedia has become "the largest bibliography in human history."

1. Wikipedia is accurate and reliable

Exactly how accurate is Wikipedia really? The website was launched in 2001. By the time it completed its third birthday, an IBM study found that "vandalism is usually repaired extremely quickly—so quickly that most users will never see its effects," and that Wikipedia had "surprisingly effective self-healing capabilities."

Before turning five years old, a blind study in the journal *Nature* searched for serious errors on a range of medical and scientific articles across the fields of pathology, toxicology, oncology, pharmaceuticals, and psychiatry. The research concluded that Wikipedia had no more serious errors than did the *Encyclopædia Britannica*.

In 2007, the German *Magazin für Computertechnik* conducted a separate study that reached the same conclusions: Wikipedia did not have any more errors in its texts than did three commercial competitors, *Bertelsmann Enzyklopädie*, *Brockhaus Multimedial* premium, and *Encarta*.[1]

By 2013, Wikipedia had become in all likelihood the most viewed medical resource in the world, with 155k articles written in over 255 languages, and 4.88 billion page views that year. Between 50-70 percent of physicians, and over 90 percent of medical students, use Wikipedia as a source for health information. Moreover, medical students using Wikipedia to prepare for a test, which was similar to the Canadian medical licensing examination, outperformed control groups using both digital textbooks and a subscription medical-reference service commonly used by doctors.

Where the research does find that Wikipedia sometimes does fall short is not in accuracy, but rather in completeness with respect to more specialized and in-depth sources. For example, a 2014 study in the scientific journal *PLOS ONE* found that, in comparison with a pharmacology textbook, Wikipedia's information on this topic was 99.7 percent accurate, but 83.8 percent complete. The authors concluded that "Wikipedia is an accurate and comprehensive source of drug-related information for undergraduate medical education."

Since then, the encyclopedia has continued to grow more robust and sophisticated. Thanks to its founder Jake Orlowitz, The Wikipedia Library now provides editors with 100,000 high-quality academic journals, free of charge, so they have access to "a sizable portion of the world's scholarship."

Today, As Orlowitz has noted, Wikipedia is cited in federal court documents, relied upon by Apple's Siri and Amazon's Alexa, and often appears as a link on the first page of Google searches. In addition, Google's Knowledge Graph, which comprises 500 billion facts about 5 billion entities, draws heavily from Wikipedia. Such content provides

[1] Dorothee Wiegand: "Entdeckungsreise. Digitale Enzyklopädien erklären die Welt." c't 6/2007, March 5, 2007, p. 136-145.

excerpts for Google search's popular Knowledge Panel, as well as capabilities to Google Assistant and Google Home.

2. Wikipedians are neither employees nor contractors

Other common misconceptions are that Wikipedia is a centralized media organization and that their employees edit content. In fact, the Wikimedia Foundation employs only about 350 people, who do not edit content as part of their job.

Rather, content is amassed, referenced, edited, and adjudicated by a large number of non-employee, non-contractor, volunteer editors. Over the past year, Wikipedia has received 259 billion pages views (up 34 percent year over year), 532 million edits (up 37 percent), of which 282 million were human user edits (up 18 percent). The last 12 months saw 4.2 million new registered users (up 27 percent). The number of editors of the English-language Wikipedia in just May 2020 was 443k, with 45k active editors who made 5 or more edits per month.

3. Wikipedia is fast

The word "wiki" comes from the Hawaiian language and literally means "quick." You can see how rapidly the edits clock is increasing here. The 4.3 billion edits are accruing approximately 700 new edits per minute.

New facts in the world that meet Wikipedia's notability standards update on the website very quickly. Though the encyclopedia does not add new knowledge within the milliseconds of "real time," living articles certainly reflect significant status changes in the world quickly enough for fact-checking applications. The more important the new information is, the faster it appears in text.

If you read a mainstream newspaper and find a story breaking above the fold today, there is an excellent chance that the new information is already on Wikipedia, with a succinct summary supported by references to verifiable facts. For example, on June 15, 2020 CNN published a story about Justice Neil Gorsuch's opinion on the Supreme Court's ruling regarding protections for LGBTQ workers under the Civil Rights Act. Before the end of the day, over 100 edits had occurred on Bostick vs. Clayton County, Georgia, with over 10 new substantiating sources.

4. Wikipedia is very much a technology

It is important to highlight some of the less visible components of the Wikipedia technology and processes – beyond simply crowdsourcing – that generate such accurate and fast results at scale. To zoom in just a little bit, while still staying at the proverbial

30,000 feet, here is a partial list of the platform's less obvious yet critically important components:

> **Bots:** Wikipedia has a Bot Approvals Group and a Bots Policy, which govern the use of automated editing. Bots can make thousands of edits per minute, including to reverse attempted vandalism. For example, when an anonymous user tried to write the word "penis" on an article about National Supreme Courts, a bot was intelligent enough to remove it instantaneously. Reversing vandalism is the remit of ClueBot NG. In addition, ORES, an API that provides machine learning as a service to Wikimedia projects, provides scores for edit quality and article quality.
>
> **Rules:** Wikipedia provides a simple set of core content policies. Information must be (1) written from a neutral point of view and (2) attributed to a reliable published source with a reputation for fact checking and accuracy. Also, (3) original research is prohibited. At the same time, there are hundreds of pages of policies and guidelines, which have developed into a veritable body of common law.
>
> **Roles and hierarchy:** Although anyone can submit an edit, Wikipedia has a formal hierarchy of administration. For most users, their number of edits, as well as the reputation of their pseudonymous personas on Wikipedia, are important to attaining status. Some especially high-level and well-defined roles include bureaucrat, administrator, steward, and arbitrator, each with specific rights. Stewards, for example, are elected annually and must have at least 80 percent approval.
>
> **Consensus and conflict:** Editors strive to reach consensus whenever possible through discussion on article and user talk pages. However, the system provides a number of different conflict-resolution mechanisms, including requests for Third Opinions, community comments, subject-specific help, editor assistance, the use of a Dispute Resolution Noticeboard, and, as a last step, passing to an Arbitration Committee.
>
> **Transparency:** Articles provide public logs of activity, in history, talk, and statistics pages.
>
> **Enforcement:** Problematic editors can be blocked for different periods of time, and the worst offenders can be banned. There are 11 major mechanisms through which pages can be locked or protected.

The phenomenal success of Wikipedia as a web platform inheres in its well-evolved and tested mechanisms and processes. The magic emerges from innumerable distributed actions that occur amid a tension between "paradoxical" elements: the system has been described as being "at once egalitarian and hierarchical, rule bound and consensus driven, collaborative and conflict driven." The encyclopedia's mechanisms are simple enough, on

one level, to be used by millions of editors; and yet they're complex enough, at another level, to require management by a smaller number of specialized technicians and jurists.

5. Wikipedia compared with journalism

By virtue of these extraordinary processes, Wikipedia has indisputably surpassed journalism as a means to determine the factuality of content. Interestingly, Wikipedia's ability to distinguish information from misinformation has also increased over the same timeframe that the ability of many other media to do the same has decreased.

The Internet has not merely lowered barriers to entry for countless new media; it has *entirely removed* them. Such outlets arguably extend down to every person on earth with a social media account. This media explosion has greatly devalued content and increased competition, even among traditional media incumbents, over narrower ideological niches. Both mainstream outlets and propaganda mills increasingly cater to people's biases, all too often in opposition to facts. Under economic pressure, editorial oversight has eroded. With regard to topics like COVID-19, misinformation can have especially harmful consequences.

Wikipedia, in contrast with any one news source, centralizes far more information in a single, searchable, and well-structured location. Rather than pandering to biases, Wikipedia systematically filters out opinions, non-factual assertions, unsubstantiated claims, fake news, and non-neutral points of view.

Wikipedia, as a process, does not operate on the ethic of a Fairness Doctrine. That is, Wikipedia does not – as many journalists still do today – reflexively give equal weight to different opinions on a controversial topic. Rather, it synthesizes many reliable sources and passes them through fact-verification mechanisms. The difference in the output is absolutely eye opening. If you visit the article titled "Global warming controversy," for example, the article summary at the top reads:

> "In the scientific literature, there is a strong consensus that global surface temperatures have increased in recent decades and that the trend is caused by human-induced emissions of greenhouse gases.[1][2][3][4][5][6] No scientific body of national or international standing disagrees with this view,[7] though a few organizations with members in extractive industries hold non-committal positions,[8] and some have attempted to convince the public that climate change isn't happening, or if the climate is changing it isn't because of human influence,[9] attempting to sow doubt in the scientific consensus.[10]
>
> The controversy is, by now, political rather than scientific: there is a scientific consensus that global warming is happening and is caused by human activity.[11] Disputes over the key scientific facts of global warming are more prevalent in the

> media than in the scientific literature, where such issues are treated as resolved, and such disputes are more prevalent in the United States than globally.[12][13]"

The article then provides a concise summary of the important scientific, political, and legal facets of the topic. This content includes a dozen quantitative charts and over 300 references to peer-reviewed journal articles and media outlets with a reputation for reliability and fact checking.

Similarly, the summary of the article titled "Vaccines and autism" is succinct and accurate:

> "Extensive investigation into vaccines and autism[1] has shown that there is no relationship between the two, causal or otherwise,[1][2][3] and that vaccine ingredients do not cause autism.[4] Vaccinologist Peter Hotez researched the growth of the false claim and concluded that its spread originated with Andrew Wakefield's fraudulent 1998 paper, with no prior paper supporting a link.[5]
>
> Despite the scientific consensus for the absence of a relationship,[1][2] the retracted paper and the anti-vaccination movement at large continue to promote myths, conspiracy theories, and misinformation linking the two.[6]"

In contrast, searching on YouTube for "vaccines and autism" returns top hits such as "Controversial researcher claims link between vaccine and autism" and "Former Congressman: Vaccines linked to autism." In this second video, former Representative Dan Burton spends half of the footage asserting that vaccines cause autism, while the newscaster spends the other half arguing that there is no connection whatsoever – both of course claiming that the facts are on their side.

The Burton video is interesting for two reasons. First, it follows the same pattern as the other video, in that different authorities are handpicked to disagree, each one providing conflicting and confusing statements about a contentious topic (albeit a topic that is scientifically uncontroversial). Second, the Burton clip is five years old; yet it rises to the top of hundreds of thousands of results.

The reason why the old Burton video floats to the very top of today's search illustrates an important difference between journalism and social media. Journalism covers current events that are of broad interest and/or (but almost always "and") because they are controversial with respect to public opinion. In an era with an exploding supply of outlets, and fierce competition to capture attention, the importance of the controversial element for journalism has dramatically increased.

When a popularity-maximizing algorithm is applied to topics that are inherently and increasingly controversial, the result is that the very most controversial and extreme examples rise to the top – even if they're five years old. In the case of the other video, the search results went all the way to Australia to pull up the clip on "Controversial

researcher" Andrew Wakefield, the discredited ex-physician, disbarred a decade ago for a famously fraudulent paper. So a popularity-maximizing algorithm can promote content that is the most controversial among an uninformed public, even when there is no scientific controversy. Worse still, the algorithm will travel quite far in space and time to surface the most conflictive content.

In the great majority of cases, promoting popular, controversial, and or attention-worthy content on social media feeds is perfectly fine and even beneficial. There are many millions of phenomenal videos showing people pushing the limits of sports, creative thought, science, entertainment, education, and many other areas. But in the specific case of videos about immunization, 32 percent of them on YouTube oppose vaccination. This trend is arguably linked to emergency-level measles outbreaks in countries like the US, the Philippines, Ukraine, Venezuela, Brazil, Italy, France, and Japan. The implications are clear: if fewer impressions of anti-vaccination content – which includes dangerous medical misinformation from the likes of *Clueless*'s Alicia Silverstone and *Deuce Bigalow*'s Rob Schneider – are seen by suggestible parents, the fewer children will die from highly contagious and easily preventable infectious disease.

While the challenge at hand may seem daunting, there are quite feasible ways to significantly mitigate harmful misinformation on social media, and certainly without throwing away more than a couple drops of the bathwater – let alone so much as rocking the baby. The next section below proposes a novel approach.

IV. AN ADAPTED SOLUTION

Social media platforms have uploaded prodigious amounts of old and newly inspired content and distributed it by personalized popularity. A tiny percentage of total content contains misinformation deleterious to public health, and a very small portion of this tiny percentage is overrepresented in the short tail of high popularity.

In parallel, Wikipedia has, by all metrics, succeeded in accurately organizing the world's information by reliably documented facts. As Clay Shirky, NYU Vice Provost for Educational Technologies, has pointed out, the free encyclopedia's success signifies no less than "a change in the nature of authority." Rather than deriving authority from an author, an institution, or a traditional media organization, "what Wikipedia suggests is that you can vest authority in a visible process." We have yet to fully recognize the value and potential of this process.

It is now upon us in the tech industry to innovate adaptive solutions that strengthen the bonds of complementarity between these different information superpowers. In the same way nations mutually benefit through more efficient trade, platforms can do the same. The basic exchange routes have already been established: Over 70 percent of traffic to Wikipedia comes from search; in the other direction, Google's fact-based Knowledge Graph and Knowledge Panel draw a significant amount of content from Wikipedia.

We must now adapt the open-source fact-checking mechanisms innovated and evolved within Wikipedia, to mitigate the hazards of misinformation on social media. It is important that the cure be feasible, non-invasive, and that it minimize side effects. A viable solution should adhere to the following constraints: there should be no upstream change to core popularity-maximizing algorithms, no change to a platform's business model, and it should allow up to 99.9 percent of revenue and content to remain untouched. These requirements will help make a proposed solution economically and politically viable, and the last one in particular greatly increases technical feasibility.

Consider, for instance, a platform like YouTube, onto which hundreds of hours of new video content are uploaded per minute. What would a viable data flow look like? To start with, perhaps after due research YouTube executives decide to implement an experimental test to target and select an extremely tiny percentage of content for external, distributed, non-employee review. The initial selection process is a matter of policy, and it occurs within the closed part of the system, such that YouTube retains complete control over which tiny proportion of videos should be eligible for external fact checking (this fact-checking eligibility selection, incidentally, would be an ideal and proactive function for Facebook's new Oversight Board).

A first proof of concept may have an extremely narrow focus. A recent study provides a good potential point of departure: research has shown that people who receive news from social media platforms are not only more likely to believe in conspiracy theories; they're also more likely to break lockdown rules. Specifically, "Some 60% of those who believe that COVID-19 symptoms were linked to 5G radiation said that much of their information on the virus came from YouTube – while of those who believed that was false, just 14% said they depended on the site. People who had ignored official advice and gone outside despite having symptoms of the virus were also far more likely to have relied on YouTube for information."

Let us imagine, then, that the only videos selected internally for an initial proof of concept would contain the terms "COVID-19", "5G", and other features most predictive of content known to espouse the false conspiracy theory that this new cellular network standard causes the disease. The eligible content could be further filtered down by a minimum impression count or acceleration of engagement, until there is a manageable list.

Next, this ranked list would pass to a new, open part of the system, a distributed review process by users with no affiliation with the company. At first, this open system would be limited to a small scale and would produce results only for internal testing. That is, results would not yet impact any non-participating user experience.

The open portion could initially comprise a relatively small number of minimally ranking Wikipedians, and later scale up to a new class of self-selected "YouTubians" passionate about the reliability of information. They would start by using, to the extent possible, the

same open source software, mechanisms, and safeguards that have successfully evolved on Wikipedia, to enable the collaborative adjudication of verifiability. That is, authority would be vested in Shirky's "visible process," which would take place on an independent MediaWiki governed by one or more social media platforms.

The facts themselves – the ground truth – would be English-language Wikipedia text itself (which is supported, as always, by bibliographic references to fact-checked sources).[2] The fact-checking platform could require that this ground truth be referenced only from articles that have minimum authorship and editorship statistics; thus, only high-visibility, central pages that have run information through the necessary processes would be used. Alternatively, social media platforms could select a small number of eligible ground-truth articles.

The designated article for fact checking, in this case, would be an article titled "Misinformation related to the COVID-19 pandemic." As this sentence is written, this page is substantiated by 634 approved references, approximately 20 percent of which appeared over the past two months. The article is secured by a Semi-Protection lock, which means that it cannot be altered by unregistered or unconfirmed users, whose accounts lack a minimum age and edit number. The page statistics show that 355 authors and 543 editors have contributed to the article.

One of the videos that may be flagged for external, crowd-sourced fact checking may be "Ebro Addresses Keri Hilson's 5G Coronavirus Conspiracy Theory." This clip features the American radio personality Ebro Darden on radio station WQHT in New York and has been viewed over 458k times since March 16th. In the video, Ebro says:

> ***Ebro:*** *You know there were people, pundits, the Internet, before 5G was rolling out, saying that 'Don't roll it out. Don't get 5G. Stop the government from 5G; it's gonna have serious health complications on the world.'*
>
> ***Co-Host:*** *I think Keri Hilson's getting dragged on the internet right now for posting all this.*
>
> ***Ebro:*** *Yes she did. Keri, go off, tell your truth, mama! There's nothing wrong with thinking about it, whether it's factual or not, should cross your mind. These microwaves and these signals in the air could have dangerous implications, ramifications. …You know, China had 5G mad early, rolled it out aggressively at the end of 2019."*
>
> ***Co-Host:*** *You know where it started, right?*
>
> ***Ebro:*** *Where?*

[2] Because Wikipedia is not static, a given fact-checking reference would link to a particular time-stamped "version" of an article from its "history".

Co-Host: *Wuhan. …That was the original city of the roll out.*

Having watched this video, a Wikipedian/YouToubian would be able to easily locate the sub-section covering 5G on the COVID-19 misinformation Wikipedia page. Then, in the simplest iteration, each reviewer would indicate that the video does in fact contain misinformation, with a reference linking to this section of the article. Any collaborative discussion would be logged on a transparent talk page, just as in Wikipedia.

In a slightly more developed iteration, a fact-check box could appear during the factually incorrect segments of the video. This adaptation to the medium of video would be especially useful for the CNN video mentioned above in which Anderson Cooper argues with former Congressman Dan Burton, because half of this video is factual and the other half is easily labeled as misinformation.

The higher the misinformation metrics and the higher the view count, the more problematic the video. Having more granular data on misinformation and reliable metrics in place, contributed through an authoritative and proven process by non-employee users, and linked to the world's largest centralized fact-checking bibliography, will be a key milestone. YouTube, or other social media platforms, could back test and forward test this system for accuracy before making any decision on how to best apply the results such that they have a well-controlled, real-world positive impact.

One of the most interesting questions for discussion is what exactly should be done with videos determined to contain different levels of harmful misinformation. We could start with the case of videos like Ebro's, which advances Kerry Hilson's 5G COVID-19 conspiracy theory, and which may very well have contributed to misunderstanding, confusion, and increased mortality rates in New York, as research suggests. A video like this one should, at the very least, have fact-check warnings linked to Wikipedia. Moreover, it should not appear anywhere near the top pages of a generic search on "COVID-19 + 5G", and it should also be demonetized. That is, if the video stays on YouTube at all, it should be caught earlier and not reach half a million views during the critical onset of the first pandemic wave in New York.

Labeling, demonetizing, and *especially* deranking may be sufficient. However, there is a strong argument for entirely removing the Ebro video and others like it, because leaving up harmful misinformation with fact-check labels is still quite problematic. Large bodies of research show that (1) people are suggestible, and (2) human psychology is guided by biased predispositions and motivated cognition. The first characteristic explains why advertising is a multibillion-dollar industry (and why governments concern themselves with monitoring public opinion and engaging in or counteracting intentional disinformation). It is in this territory of suggestibility in which elections are fought; and this is the location of the 31 percent of the population that is not yet sure if they'd accept a COVID-19 vaccine.

The second human-nature constraint explains why people continue to support political positions, or endorse conspiracy theories, even in the face of factual information to the contrary. A property of conspiracy theories is that it is nearly impossible to falsify them in the perception of their believers. Strong political biases are only somewhat more malleable by comparison. Therefore, major social media platforms can and should decide to remove a carefully determined sliver of harmful misinformation entirely in many cases. After all, this removal has already been occurring beautifully on Wikipedia, even with the encyclopedia's far more open administrative structure. Fact-check warnings alone cannot solve the infodemic. If they are used, however, it is imperative that platforms derank flagged content to minimize its distribution. In addition, it is preferable that warning labels are not placed on content by "elite" Silicon Valley tech companies, or even by centralized fact-checking media outlets, but rather in the manner proposed above, with reference to the world's largest bibliography.

There are at least two cases where misinformation should probably not be removed, but rather left up with fact-check labels. The first case concerns content posted by elected officials, candidates for public office, or government agencies. It is important that a social media platform and its fact-checking system not remove factually incorrect postings by current or potential public leaders, because people have a democratic right to judge these leaders' level of sophistication and truthfulness.

The second case entails content posted by reliable, independent, published sources with a reputation for fact checking and accuracy – that is, sources eligible to serve as bibliographic references on Wikipedia. The CNN video of Anderson Cooper arguing with former Congressman Burton, for example, should be left up with a fact-check label appearing nearly every moment the politician speaks. However, this video should not appear on the first page of the given search query. Search results can still sort by controversy – just not when videos (1) pertain to a topic eligible for external fact checking, like vaccinations, and (2) incur significant user-generated misinformation scores, and (3) would surpass a minimum impression threshold.

In sum, Keri Hilson, Alicia Silverstone, and Rob Schneider do not have a right against a private platform to spread dangerous medical misinformation. However, content posted by public leaders or government agencies, as well as that appearing on media publishers with a reputation for fact checking, should be left up with fact-check labels and judged (and in some cases deranked). Much evidence from the field of game theory shows that enforcement is necessary for attaining good behavior. The absence of punishment encourages bad behavior in this context, which has been proven out empirically in the real-life social media experiment.

So what should Twitter do next time President Trump uploads factually incorrect information to their platform? They ought to leave up his tweet with a crowd-sourced fact-check warning that points directly to Wikipedia, not back to Twitter.

Here is how this would have worked earlier this year: on May 19, Trump tweeted that the World Health Organization had ignored reports by *The Lancet* medical journal of the virus spreading in Wuhan back in December 2019. Within hours, *The Lancet* denied that they had published any such reports in 2019, and this news story was covered by different trusted sources. The correct information immediately and succinctly appeared on the Wikipedia page entitled "Veracity of statements by Donald Trump," under the heading "Coronavirus pandemic." Twitter should have left the Tweet up with a fact-check label pointing precisely to the relevant paragraph at this location.

V. EXPECTED IMPACTS

The solution outlined above, or a similar one, will have a number of significant positive effects. To begin with, the world's exposure to a limited and carefully targeted set of harmful and factually incorrect content will be greatly reduced. The expected impact of this alone includes greater compliance with disease-control policies, greater immunization rates, and far fewer people dying from preventable diseases like the measles and COVID-19, to name only a few achievable outcomes.

The proposed system will also improve over time. As more and more content elicits granular fact-check data provided by crowd-sourced human judgment, and filtered through proven open-source processes, a virtuous feedback loop can be established to train ML algorithms to more accurately identify likely misinformation at scale.

By reducing exposure to easily falsifiable and harmful content, social media platforms will pre-empt the constant stream of criticism from conventional media outlets, which otherwise is on the rise and shows no sign of abating. The longer negative press continues, the higher the probability of government intimidation and adverse new regulation.

By having non-employee users adjudicate verifiability, albeit within an extremely narrow band of eligible content, social media platforms will have a stronger political argument against attempts to jawbone them into protecting a particular point of view (e.g., by Trump) or to carve back or even eliminate Section 230 (e.g., by Biden). If Section 230 reformers do succeed in removing immunization for third-party misinformation on specific topics like vaccinations, then it will behoove social media platforms to (1) develop a more scalable and cost-effective approach to fact checking such topics, and (2) to lobby for continued immunization for open-source fact-checking MediaWikis and the content retained by them. Avoiding power struggles between social media companies and governments is extremely important to maintaining societies that are open and free. To achieve this peace requires a pragmatic compromise mediated by Rosenbaum's "magical entity."

This approach also makes it easier for social media companies to deal with their most problematic users. Consider the case of professional conspiracy theorist Alex Jones, who

spreads misinformation about vaccines and autism, claims global warming is a hoax, and perpetrates the appalling lie that the tragic Sandy Hook shooting, in which 20 children were murdered, was faked by "crises actors." Rather than making a series of difficult and reactive internal policy decisions, each piece of eligible content could be sent for external review as each toxic posting is uploaded. If the repeat offender is not discouraged after some number of times, he or she could receive temporary suspension or a permanent ban. This way, enforcement is a function of crowd-sourced fact-checking metrics – not solely company policy. In addition, a platform's users' exposure to harmful misinformation will be greatly reduced from the outset. We should not all know who Alex Jones is.

Transferring the fact-checking function, and the metrics upon which enforcement decisions are made, outside of a social media company will also greatly reduce reputational risk in the court of public opinion. Talking heads like Tucker Carlson will no longer have a constant stream of fodder to burn in videos like this one. Here, Carlson lambasts the censors of free speech at Twitter, including the company's Head of Site Integrity, who Carlson claims is an activist who compared presidential counselor Kellyanne Conway to the Nazi Minister of Propaganda, Joseph Goebbels.

There are far too many issues with the Carlson clip to unpack here. Suffice it to say that the entire dynamic in this video should be avoided at all cost. Instead, Twitter and other social media platforms should transfer the fact-checking authority away from individual employees and over to the external, transparent, and authoritative process. Carlson should have to fight it out not with Twitter and an allegedly anti-conservative executive, but instead be forced to face the facts of Wikipedia, substantiated by its verified bibliographic references. In particular, Twitter's fact check on May 27 should have pointed to the section on the "Reliability of postal ballots" in the Wikipedia article "Postal voting in the United States." A quick check shows that, even after including all of the cases from the database of the conservative Heritage Foundation think tank, there are fewer than 800 total known cases of absentee voter fraud in the United States between the years 2000 and 2018. Twitter's fact check was correct, their decision to place a label was correct, but the way in which they did it – and all the times they do not – makes all the difference.

Social media platforms will benefit in other ways as well. A steep reduction in ad impressions on media containing misinformation will be greatly welcomed by the largest advertisers, like Procter & Gamble and their peers at the Global Alliance for Responsible Media. Also, if Facebook decides to take on even just the most high-profile political misinformation it will help win back advertisers that are now boycotting their media,

such as The North Face, Patagonia, REI, Unilever, Coca Cola, Microsoft, Starbucks, Honda, Verizon, as well as over 1,100 other brands.[3]

Finally, this approach to fact checking could eventually be applied to political advertising. Political ads could categorically be subjected to open-source fact checking (before going live), and ad campaigns running content containing misinformation could be shut down, thus improving the reliability of information in political ads that do pass through this filter[4]. Moreover, social media platforms would not have to jettison revenue from political advertising, as Twitter already did in 2019, and as Facebook has now partially done as well, by allowing users to turn off political ads. In a better world, social media companies will make money from political advertising, deceitful advertisers will be punished, and much more reliable information will reach voters.

As a potential downside, lawsuits could arise from fact-check labels, demonetization, deranking, or content removal. However, today's Section 230 would keep most such lawsuits at bay. It is important that any future Section 230 reform uphold immunity for open-source fact-checking MediaWikis and the content they pass through. In a world with crowd-sourced fact checking against Wikipedia, any such lawsuits should be easier to win for the social media platform, and more damaging for the factually incorrect party, in that they will raise awareness of the actual facts and their reliable sources.

Focusing attention on a unitary source of truth will also slightly weaken filter bubbles and political polarization, at least at their most extreme edges.

VI. NEXT STEPS

This paper is intended to serve merely as a starting point for further discussion. The adapted solution described above is a high-level concept. More research is needed to better understand the efforts undertaken to date, challenges and opportunities, and the best ways to adapt this novel approach so that it complements current infrastructure and mechanisms, with regard to both the relevant Wiki processes and social media platforms.

---

[3] It should be noted that this boycott is motivated by Facebook's decision to leave up presidential posts containing both misinformation, as well as alleged incitement to violence, according to the explicit reasons given by IPG. Advertisers and agencies are arguably singling out Facebook in large part due to this particular platform's conspicuous insistence on not fact-checking political misinformation. Four out of the 10 Recommendations given to Facebook by the boycotting coalition, Stop Hate for Profit, specify "misinformation." Recommendation 4, for example, is to remove groups focused on issues such as "violent conspiracies, Holocaust denialism, vaccine misinformation, and climate denialism" – all of which are amenable to falsification through Wikipedia-referenced fact checking. Correcting this misinformation policy, and implementing the corresponding scalable solution, is far easier than dealing with the issue of alleged presidential incitement, or with hate speech, which cannot be as easily defined as misinformation, or falsified against a pre-existing ground truth in the same way. However some of the same tools developed to combat misinformation have applications for combatting hate speech, which is also notably absent from Wikipedia articles.

[4] Removal of this political content could apply to paid ad campaigns and paid content promotion; unpaid posts or tweets by elected officials could still be left up with fact-check labels.

Research should be conducted at very small scale to begin with, to study the potential impact on the Wikipedia ecosystem. It is important to know with certainty what potential positive or negative effects may result when a small number of high-profile English-language Wikipedia articles are used as a ground truth, for an independent (social-media-governed) fact-checking MediaWiki. If the net impact is positive, then scaling up the fact-checking platform could significantly strengthen Wikipedia by minting valuable new editors, who may add content to articles in the process of fact checking.

Research on the motivation of potential fact checkers is also important, to understand the quantity and scale of effort that can power this content-moderation mechanism. From one perspective, some proportion of the relatively small group of serious Wikipedia editors would not want to work for the "benefit" of a for-profit company. On the other hand, social-media-audience fact checkers will likely comprise a far larger pool of people. And there are precedents for crowd-sourced work contributing to large tech companies. For example, Local Guides enrich Google Maps with a significant amount of information; this motivation loop works because they are not intrinsically motivated to work for Google, but rather to help friends and family.

It will be likely be important, in recruiting fact checkers, to communicate two important points: (1) their fact checking is for the benefit of the community of users, to reduce misinformation in the world; and (2) harmful content will be deranked and demonetized, reducing the profitability of bad content for both third-party creators and the tech platform.

Altruism and punishment should be enough to motivate sufficient numbers of fact checkers. Research shows that people will happily punish moral transgressions, even incurring a monetary cost to themselves, in exchange for the resulting stimulation in their caudate nucleus, which is highly innervated by dopamine neurons. Neuroscientists have linked this brain structure to reward processing in rats, nonhuman primates, and humans. "Reinforcers" such as nicotine, cocaine, and monetary rewards activate the caudate nucleus. This mechanism may account for the fact that costly arguing over moral issues already constitutes a significant proportion of user activity on social media today. The prospect of fact checkers engaging in a similar activity for "free" is therefore quite likely.

An opposite problem could arise if too many fact checkers emerge – but only if the open-source adjudication processes are not sufficiently well developed. In this scenario, more complexity, AI-powered anti-vandalism bots, hierarchy, and conflict-resolution mechanisms will be needed, as have evolved already on Wikipedia.

In addition to more research, a multidisciplinary team should be recruited to plan and implement the proof of concept. This team will of course require expertise in data engineering, data science, product management, front-end engineering and design. It will also need experts in behavioral science, law, tech policy, and government relations, as

well as Wikipedia and its technology, Wikimedia Foundation, culture, community, and software. This endeavor will require both generalist and specialist vision.

What is already clear, though, is that our top websites are complementary in the ways that they have evolved to organize the world's information. Their suitability for mutually strengthening one another further is extremely compelling.

How quickly we begin adapting the most successful fact-verification technologies to the largest popularity-driven content platforms has immensely important implications for us all. If we retain the status quo, we may live in an increasingly divisive and dangerous post-factual era. However, if we can mitigate key areas of misinformation on social media platforms anywhere near as well as already achieved on Wikipedia, then our information age will unequivocally succeed in increasing shared knowledge, understanding, health, and wellbeing.

Acknowledgements and Views Disclaimer: I would like to express my gratitude to Laurent Crenshaw, Eric Goldman, Alex Komoroske, Adam Leonard, Jake Orlowitz, Andrew Shepard, and many other anonymous experts for their valuable feedback and advice during the research and writing of this paper. The views, opinions, positions, strategies, and recommendations expressed in the text belong solely to the author, and not necessarily to the reviewers above, the author's employer, or other organizations or individuals.