The Quagmire Model: How a small step can trap you in the quagmire of misinformation, hate speech and denialism

Eduardo Camilo da Silva
Universidade Federal Fluminense (UFF)
ecamilo@id.uff.br

Claudio Henrique da Silveira Barbedo
IBMEC Business School
cbarbedo72@gmail.com

Eugenio Rangel Marins
Nuclear Engineering Institute
eugenio@ien.gov.br

Rose Marie Santini
Universidade Federal do Rio de Janeiro (UFRJ)
marie.santini@eco.ufrj.br

ABSTRACT

This study aims to explain how social media (SM) users, whilst searching for information, can be trapped into a quagmire of misinformation, even when they have no denialist inclinations or sympathy for hate groups. We analyze the interactions between cognitive biases and deep preference learning algorithms (DPL), as SM companies use DPL to curate the content conveyed to its users. The study proposes a model for users' behavior and explain how SM business model allows new information to be introduced in the quagmire and change user's opinions in a way desired by a customer willing to pay for it. The model explains why some popular tactics against misinformation, as censorship and fact-checking, achieve very poor results. We suggest that policies promoting face-to-face encounters in friendly environments can be more effective in that struggle. We believe the model can help decision makers in developing more efficient anti-misinformation policies.

Keywords:deep preference learning; disinformation; epistemic security; hate speech; social media network.

1. INTRODUCTION

Yeh *et al.* (2022) claim that social networks are extremely important in allowing personal interaction, developing commerce platforms and monitoring users' behavior. From a personal point of view, users seek networks for connection, experience sharing and, inevitably, information about events that interest themselves or the people with whom they are connected.

Social networks, however, are not a public asset created and made available for the wellness of society. They are products owned by private companies seeking profit and growth. Its business model generates revenue from the sale of this audience, *i.e.*, the time and attention of its users, to any organizations or people who are willing to pay to have their message delivered to these users.

For the attention of this audience to have value, it needs to be very numerous and that users dedicate time, attention and interact with the network's messages giving likes, posting comments or even simply by watching the videos it offers.

Thus, social networks have only one parameter for evaluating posts quality: the number of user's interactions. This goes both ways. If, on the one hand, networks benefit from these likes, on the other, users are happy to receive them. According to Empoli (2021), we are social creatures, we depend on people's approval and each like is a maternal caress on our ego. According to one of the founders of Facebook, Sean Parker, our need for recognition is the base for the entire architecture of Facebook. As reported in Lanier (2018):

> QUOTE
>
> "We need to sort of give you a little dopamine hit every once in a while, because someone liked or commented on a photo or a post or whatever... It's a social-validation feedback loop... exactly the kind of thing that a hacker like myself would come up with, because you're exploiting a vulnerability in human psychology... The inventors, creators -- it's me, it's Mark [Zuckerberg], it's Kevin Systrom on Instagram, it's all of these people -- understood this consciously. And we did it anyway...it literally changes your relationship with society, with each other... It probably interferes with productivity in weird ways. God only knows what it's doing to our children's brains."

For normal users, the social network is a source of information and brings a sense of belonging and approval by their community. What they do not realize is that in this relationship, the user is not the customer, but the product. What social media companies do is to sell their attention to whoever is willing to pay for it.

This logic, implemented by Deep Preference Learning (DPL) algorithms, may combine with users' cognitive biases, known as confirmation bias and availability bias. When this happens, an interaction with an apparently innocent post can drag the user into a quagmire of lies, hate speech and denialism. This way, disinformation, deliberately produced, may find acceptance with the user, even more so in a situation of social isolation, as is happening since 2020, due to actions to combat the COVID-19 pandemic.

This study proposes a theoretical model that explains how this vicious circle works, punctuating and explaining the role of each of the system's elements. The study aims to help understand this phenomenon, which threatens the epistemic security of our civilization and prevents us from being able to face major challenges as a society, basing our conclusions on the best scientific evidence available (Seger et al., 2020). The idea is that the understanding of this phenomenon's components guides the creation of policies and actions aimed at combating disinformation, hate speech and denialism, focused on interrupting the vicious cycle that constitutes this quagmire.

Section 2 presents and discusses the concepts that will compose each proposition of the Quagmire Model. Section 3 analyses the consequences, for users and for society, that arise from the Model, including an explanation of why some actions currently used to combat disinformation have proved ineffective. Section 4 presents conclusions and recommendations on how policies and actions should address the problem, as well as suggestions for future studies.

2. THEORETICAL REVIEW

2.1 Cognitive Biases and Heuristics

2.1.1 Availability Heuristic

The availability heuristic, described by Tversky and Kahneman (1973), is the mechanism by which the decision maker considers more likely, more important, or more frequent, the most easily remembered facts, rather than examining alternatives not so readily available in memory. They proposed the following test, with study participants, and reported this result:

> "Suppose you sample a word at random from an English text. Is it more likely that the word starts with a K, or that K is its third letter? According to our thesis, people answer such a question by comparing the availability of the two categories, i.e., by assessing the ease with which instances of the two categories come to mind. It is certainly easier to think of words that start with a K than of words where K is in the third position. If the judgment of frequency is mediated by assessed availability, then words that start with K should be judged more frequent, but in fact, a typical text contains twice as many words in which K is in the third position than words that start with K."

This result, as well as others reported in this seminal article, suggests a systematic distortion of our judgment on the importance and frequency of certain events by the availability of said events in our memory. People typically find it more likely to die in a shark attack (1 chance in 300 million) than being hit by a plane wreck (1 chance in 10 million)[1].

Riddle (2010) suggests that people who watch violent scenes on television have the perception of the social reality in which they live altered. Sjöberg and Engelberg (2010) suggest that films with high-risk scenes change participants' perception of risk and that this change has a prevalence of at least 10 days.

Exposure to social media leverages this heuristic because exposure is constant, intense, and focused, so opinions can be shaped by the volume and frequency of posts that expose the user to content coming from the same sources selected by the DPL algorithm using post-based popularity.

[1] "Odds and ends - The San Diego Union-Tribune". legacy.sandiegouniontribune.com. Archived from the original on 2019-03-25.

2.1.2 Confirmation Bias, Filter Bubbles and Echo-Chambers

Confirmation bias is a phenomenon, extensively described and validated by researchers, that explains how people, when exposed to information about a certain subject, tend to consider only information that confirms their previously established beliefs and to disregard information that confronts these beliefs (Nickerson, 1998). This effect is more intense the stronger the prior belief, before exposure to the information. Confirmation bias cannot be eliminated, but it can be minimized with critical thinking techniques.

This bias affects the attitude of the individual in the decision-making process, vitiating the search for information and the interpretation of the information obtained and, from these deviations in the process, several effects are reported: attitude polarization, when differences of opinions are increased even when both parties are exposed to the same evidence (Lord et al., 1979; Kuhn and Lao, 1996; Taber and Lodge, 2006; Dahlgren, 2020); belief perseverance, when individuals maintain their belief even in the face of contrary evidence (Ross and Anderson, 1982; Tversky and Kahneman, 1974; Nickerson, 1998); the irrational primacy effect, which refers to the individual's attachment to the first opinion found in a series (Baron, 2000) and illusory correlation, when the person finds false correlations between events (Plou, 1993; Fine, 2006).

We call Echo-Chambers the environments created by the continuous selection of information that reinforces previous opinions, often directed by a source interested in the intellectual isolation of individuals and their alignment around certain opinions. The creation of Echo-Chambers exploits confirmation bias so that people do not realize they are being directed to opinions chosen by other sources, which can include hate speech and conspiracy theories dangerous to society, such as preaching against vaccines or groups specific ethnicities.

Social media environments, curated by algorithms that select the posts to which users are exposed, lead to the creation of Bubble Filters, which, in turn, reinforce confirmation bias,

creating a spiral of reinforcement of previous beliefs, even in the face of evidence that contradicts them (Spohr, 2017).

Bubble Filter is a state of intellectual isolation that can be caused by searches for information on websites or social networks, curated by preference learning algorithms, which select results based on the user's previous searches or interactions with posts, ads or links clicked in the past (Bozdag, 2013).

The term Bubble Filter was created by Pariser (2011) and explains how curative algorithms can be used to radicalize beliefs and isolate people in opinions that are difficult to change by traditional means of convincing, such as presenting evidence contrary to these positions and fact-checking.

2.1.3 Deep Preference Learning Curator Algorithms

As the business model of social networks depends on audience growth and the engagement of network users, these networks use curator algorithms that, based on deep learning techniques, seek to learn users' preferences because it is known, from the bias of confirmation, that posts close to these preferences are more likely to get some form of engagement, such as likes, comments, or attention, as measured by how long the user looks at the post before "scrolling" the timeline.

Datta (2021) discussed how an agent's inherent worldview influences the software's embedded logic, creating a fundamentally biased design paradox. According to the author, the intention of social media DPL algorithms to maximize user interactions with the network directs the results to the exposure of confirming news of users' previous opinions, creating echo chambers, so that the confirmation bias is amplified using filter bubbles, or "algorithmic editing", which display to individuals only information they are likely to agree with, while excluding opposing views.

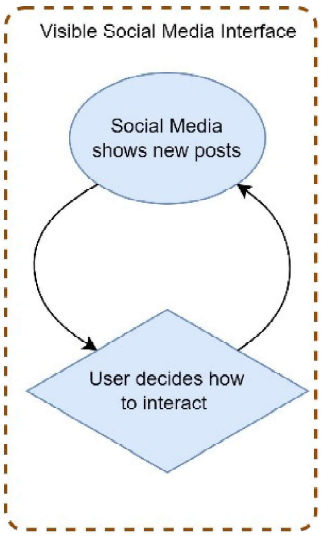
The power of these algorithms can alter significantly users' behavior, so Lanier (2018) decided to call them BUMMER (Behaviors of User Modified and Made into an Empire for Rent). This ability is directly linked to the interaction of curative algorithms with behavioral biases. The biggest problem is that the revenue of social media companies comes primarily from customers who pay for their messages to be taken to the users they choose, and the algorithms they also serve this model. So, the attention and, consequently, the behavior of users are rented to these customers.

3. THE QUAGMIRE MODEL

We introduce a social misinformation diffusion model based on confirmation bias and availability bias. The model explains how a person with only one common opinion with a denialist ends up being dragged into the quagmire of denialism and hate speech.

We started in an information sharing economy where social media does the work of providing interesting news, rumors, conspiracy theories and even noise, to encourage users' interaction.

Figure 1. Visible Social Media interface

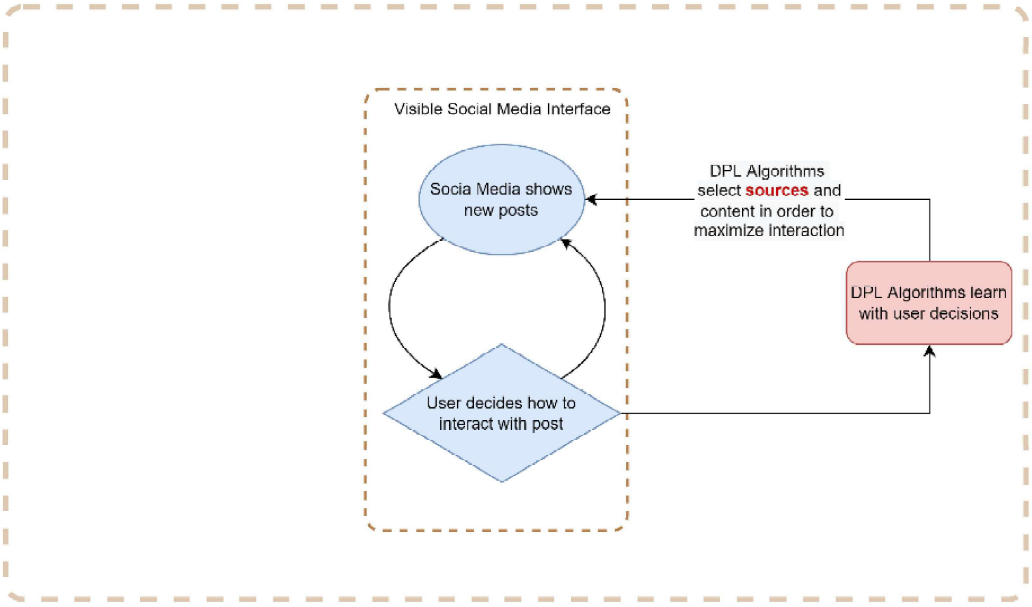


Social media platforms select customized sources to improve website traffic, increase conversion rates, and propose a variety of affordances as "Like," or "Dislike," or "Subscribe" to create a user preference database. This database is built by DPL algorithms that filter data to

match individuals' preference and learn with day-to-day choices. Thus, we argue that social media platforms generate content that promotes attention and encourages users to share it with their social networks.

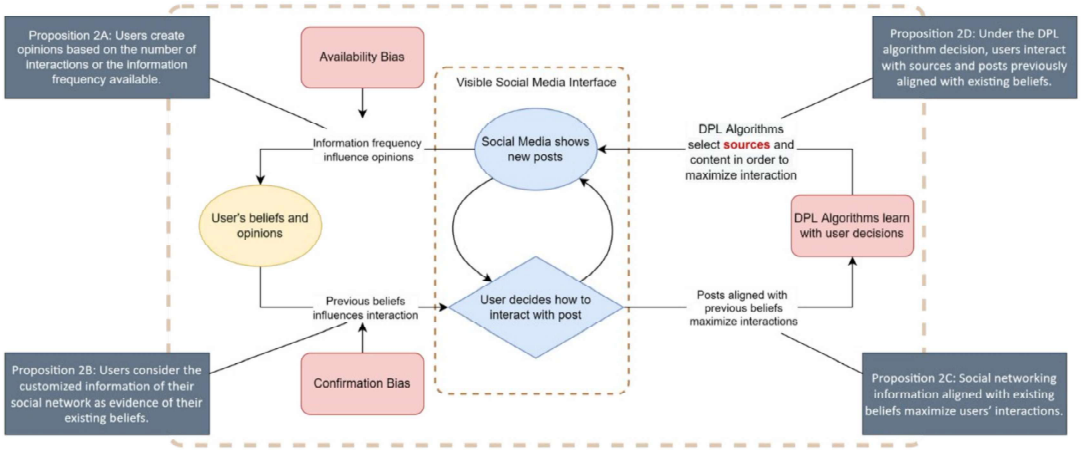
Proposition 1: Social media platforms maximize interactions via the choices of customized sources.

Figure 2. Proposition 1: Social media platforms maximize interactions via the choices of customized sources and content.



Availability bias. When the DPL algorithm focus on contents that a user will probably want to see, it will increase the user´s perception of truth about said contents, thus causing judgment distortions, i.e., the information's importance is based on which one is the most available rather than which one is the most representative. This is the definition of availability bias or availability heuristic. It means that users' decision-making process is affected by their most accessible information. Luo and Markowitz (2020) state that the amount of 'likes' boosts the supposed credibility of real headlines and of fake headlines.

Figure 3. Proposition 2A: Users create opinions based on the number of interactions or on the availability of the information; Proposition 2B: Users consider the customized information of their social network as evidence of their existing beliefs; Proposition 2C: Social networking information, aligned with existing beliefs, maximize users' interactions and Proposition 2D: Under the DPL algorithm decision, users interact with sources and posts previously aligned with existing beliefs.



Proposition 2A: Users create opinions based on the number of interactions or on the availability of the information

The passive act of searching for information and adding it to their own social media is now an active act of spreading their own opinions and influencing the formation of other individuals' opinions (Frees and Koch, 2018). Users create opinions based on the interactions and information available in their own social media. This causes another judgment distortion by only processing information that support their own beliefs. This is the definition of confirmation bias. In the social media context, there is a tendency to interpret the social network's customized information as evidence of one's existing beliefs.

Confirmation bias is an efficient way to process information when overwhelmed with social media information and cannot process each post carefully.

Proposition 2B: Users consider the social network's customized information as evidence of their existing beliefs.

As social media customized information endorses the opinion of individuals, it motivates users to express their opinion, also via social media. A shared content from user to user can form large cascades of resharings. Cheng et al (2014) state that cascades are an information-sharing mechanism where content reaches social media users. They occur when individuals in a population exhibit herd-like behavior. Content sharing has thus become a crucial information discovery tool in social networking web sites (Cheng et al., 2014).

Proposition 2C: Social networking information aligned with existing beliefs maximize users' interactions.

Now we return to the DPL algorithm learning process to feed the looping procedure. The DPL aim is to decide about what users want to see on the platform and to present posts in the best order the algorithm decides. Besides, users are more likely to interact with content that is concordant with their previous beliefs. According to Moravec et al (2019), users are more likely to rely on belief consistent news.

Ruffo *et al.* (2021) state repetition is a well-known propaganda mechanism exploited by social media. It allows building a reinforcement feedback loop on the user's online feed.

Proposition 2D: Under the DPL algorithm decision, users interact with sources and posts previously aligned with existing beliefs.

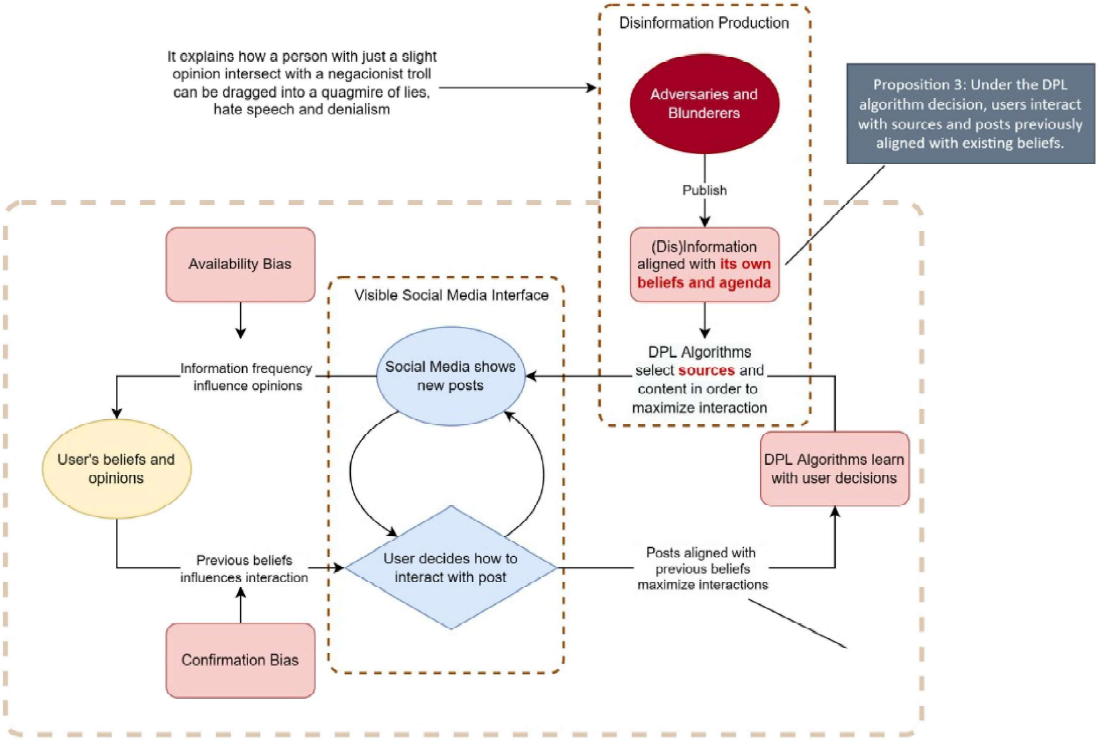
The next step is when a fake news publisher or a troll intends to make false claims for personal purpose.

When a fake news publisher posts a polarized opinion, the social media algorithm creates a filter bubble (Pariser, 2011), *i.e.*, a content that the platforms will apply on some user's online feed. In addition, this content is shared by a user's connection, so this process triggers the echo chambers (Sunstein, 2002). Ruffo et al. (2021) define echo chambers as tightly knit clusters of individuals that stay interacting until they get radicalized because of a reinforcing feedback

loop. Therefore, there is a rise of misinformation and skeptical views about social issues supported by one of the most effective techniques of persuasion for online news, the repetition.

Proposition 3: The fake news publisher posts misinformation selected from sources previously aligned with existing beliefs that will be selected by DPL algorithms to be presented to the user.

Figure 4. Proposition 3: The fake news publisher posts misinformation selected from sources previously aligned with existing beliefs that will be selected by DPL algorithms to be presented to the user.



Selected sources and posts feed the social media network and posts numbers influence opinions and decisions of others. In a conventional society, face to face interaction allows individuals to consider opposing views in a conversation. It allows for a better exchange of information and establishes trust between people. However, adjustments in the way people live, as lockdowns, shelter in place or social distancing policies, thwart empathy effects.

In a lockdown environment, given the lack of a consensus about an issue, people turn to informal sources of information to share their thoughts and to discuss issues. At the same time, social media news reports potential misleading stories entangled with social media posts feedback. It inhibits the understanding of different or opposing opinions.

Allamong and Peterson (2021) show that empathic ability may play a key role in changing people's behavior and that empathy depends upon the respondent's partisanship, the target's partisanship, and the interaction among these persons. Yet according to them, polarization reduces empathy.

So, an isolated environment can reduce opportunities to create empathy that would otherwise be created by the interaction among people. It also prevents serendipity to expose a user to new ideas with polarization reducing potential.

This results in a situation that reinforces the bubble filters making people more vulnerable to radical opinions, like conspiracy theories and hate speech.

4. CONSEQUENCES OF THE MODEL

The kernel of the Quagmire Model is in the motivation of social networks to select posts that primarily aim to maximize interactions such as clicks, likes, comments, etc.

The Social Media Companies BUSINESS MODEL is based on:

- Capturing users' attention by keeping them connected to the NETWORK.
- Extract as much information as possible about the user's preferences.
- Organize the preferences, habits and characteristics of each user with machine learning algorithms.
- Select posts through curation algorithms that are based on these preferences.
- Sell access to that audience to anyone willing to pay.
- Expose users to content produced by audience buyers.

As audience buyers can be the producers of disinformation, people are completely exposed to contents they produce.

It turns out that availability and confirmation cognitive biases deepen convictions and, ultimately, drive user's behavior according to chosen posts. These biases are universal and, for people to avoid their effects, they need favorable and prepared environments to protect them from exposure to situations that trigger said biases.

To avoid availability bias, people must have access to varied and credible information so that they are not induced to believe specific information just because it is frequently being repeated.

In the case of confirmation bias, they need to be exposed to diverse information in an environment that encourages them to judge conflicting information based on the merits of its contents, without contamination by their prior beliefs. Pariser (2011) suggests that people need randomization and serendipity.

In recent years, the social isolation necessary to combat COVID-19 has reduced personal interactions and, consequently, opportunities for developing empathy between people with different opinions, which has contributed to further deepening the quagmire of misinformation people are living in.

The Quagmire Model explains why censorship and fact checking do not work effectively enough to protect our epistemic security and thus prevent the growth of disinformation misuse.

Censorship seeks to prevent sources that produce disinformation content from disseminating this material on the networks. That is, it tries to prevent this content from being presented to users defined to receive them by the DPL algorithms.

The fact check, in turn, aims to expose the falsity of the content in order to discourage the user from believing in it or its sources.

Three factors drastically reduce the effectiveness of these devices. The speed of networks, the temporality of actions and the strength of bubble filters.

In social networks, the content propagation speed is very high. Mitigation actions, in turn, occur only after the release of the content to be censored or unmasked. The combination of these two factors means that the content has already taken effect on many users when control devices kick in.

In addition, among other effects, the conspiratorial opinions reinforcement, typical of bubble filters, that information coming from opponents of the bubble members is malicious and fallacious, greatly reduces the effect that these devices were intended to have.

For a person outside the quagmire to have any effect on another person trapped in the quagmire, the vicious cycle of misinformation and bias needs to be broken by creating situations in which the disinformation produced by chaos engineers is no longer well received, but, on the contrary, causes a cognitive dissonance that makes individuals allow themselves to be exposed to new information.

From there, even if the person keeps getting information via social networks, it may be that the algorithms themselves that are in charge of presenting new information that reinforces this new position may burst the filter bubbles and reduce or even eliminate the cognitive biases effects.

5. FINAL CONSIDERATIONS

This model explains how a person with only one opinion in common with a troll ends up being dragged into the quagmire of polarization and hate speech.

According to the model, it is possible that many people are immersed and active in a parallel reality, believing in conspiracy theories, supporting discriminatory leaders and initiatives and reverberating hate speech without having consciously chosen to be in such a situation.

What may have started as a simple innocent search, driven by an initial belief, may have escalated to gradual exposure to malicious posts such that, like a person falling into a quagmire, movement within it only makes the person more and more entrapped.

The only way to get individuals out of the quagmire is with outside help. Someone must pull them out of there and bring them ashore.

In the quagmire of disinformation, help needs to work by breaking the chains of the vicious cycle. There is little point in trying to combat cognitive biases directly. These biases are very strong and will always act when conditions allow.

This article is not intended to exhaust these alternatives, but it does provide a framework for policymakers, activists, practitioners and academics to base their actions and projects on combating disinformation, hate speech and denialism.

The biggest risk remains on the business model of the companies that control social networks, because selling the audience could cause the disinformation produced by the engineers of chaos to be delivered to users, even if it means going against current preferences.

The most effective way to break this cycle is through face-to-face interaction with people that think differently, but for whom the user trapped in the quagmire has empathy. The collision of the person's affective vision with the received misinformation can cause the cognitive dissonance necessary to make this user amenable to information contrary to what is being exposed in the quagmire, bringing the serendipity and randomization necessaries to help one to find the strength to get out of it.

Other tactics and strategies aimed at getting people out of the quagmire can and should be pursued and developed. The purpose of this article is to provide a framework for thinking about creating and implementing these policies.

REFERENCES

Allamong, M. B., & Peterson, D. A. (2020) Screw Those Guys: Polarization, Empathy, and Attitudes About Out-Partisans. *Political Psychology*, 42(3): 349–369. DOI: 10.1111/pops.12701.

Baron, J. (2000) *Thinking and Deciding*. 3rd ed. New York: Cambridge University Press.

Bozdag, E. (2013) Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, 15(3): 209–227. DOI: 10.1007/s10676-013-9321-6.

Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., & Leskovec, J. (2014). *Can cascades be predicted?*. In Proceedings of the 23rd international conference on World wide web (pp. 925-936).

Da Empoli, G. (2019) *Os engenheiros do caos: Como as fake news, as teorias da conspiração e os algoritmos estão sendo utilizados para disseminar ódio, medo e influenciar eleições*. Vestígio Editora.

Dahlgren, P. (2020) *Media Echo Chambers: Selective Exposure and Confirmation Bias in Media Use, and Its Consequences for Political Polarization*. Gothenburg: University of Gothenburg.

Datta, P., Whitmore, M., & Nwankpa, J. K. (2021). A perfect storm: social media news, psychological biases, and AI. *Digital Threats: Research and Practice*, 2(2), 1-21.

Fine, C. (2006) *A Mind of Its Own: How Your Brain Distorts and Deceives*. Cambridge UK: Icon Books.

Frees, B., & Koch, W. (2018). ARD/ZDF-onlinestudie 2018: zuwachs bei medialer internetnutzung und kommunikation. *Media Perspektiven*, 9, 398–413.

Kuhn, D., & Lao, J. (1996) Effects of Evidence on Attitudes: Is Polarization the Norm?: *Psychological Science*, 7(2): 115–120. DOI: 10.1111/j.1467-9280.1996.tb00340.x.

Lanier J (2018) *Ten Arguments for Deleting Your Social Media Accounts Right Now*. New York: Henry Holt and Company.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11), 2098.

Luo, M., Hancock, J. T., & Markowitz, D. M. (2022). Credibility perceptions and detection accuracy of fake news headlines on social media: Effects of truth-bias and endorsement cues. *Communication Research*, 49(2), 171-195.

Moravec, P., Minas, R., & Dennis, A. R. (2018). *Fake news on social media: People believe what they want to believe when it makes no sense at all*. Kelley School of Business research paper, (18-87).

Nickerson, R. S. (1998) Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2(2): 175–220. DOI: 10.1037/1089-2680.2.2.175.

The Filter Bubble: *What the Internet Is Hiding from You*. penguin UK.

Plous, S. (1993). *The psychology of judgment and decision making*. Mcgraw-Hill Book Company.

Riddle, K. (2010) Always on My Mind: Exploring How Frequent, Recent, and Vivid Television Portrayals Are Used in the Formation of Social Reality Judgments. *Media Psychology*, 13(2): 155–179. DOI: 10.1080/15213261003800140.

Ross, L., & Anderson, C. A. (1982) *Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments*. 129–152. DOI: 10.1017/cbo9780511809477.010.

Ruffo, G., Semeraro, A., Giachanou, A., & Rosso, P. (2021). *Surveying the research on fake news in social media: a tale of networks and language*. CoRR, abs/2109.07909.

Seger, E., Avin, S., Pearson, G., Briers, M., Heigeartaigh, S. Ó., Bacon, H., ... & Weller, A. (2020). *Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world.*

Sjöberg, L., & Engelberg, E. (2010) Risk Perception and Movies: A Study of Availability as a Factor in Risk Perception. *Risk Analysis*, 30(1): 95–106. DOI: 10.1111/j.1539-6924.2009.01335.x.

Spohr, D. (2017) Fake news and ideological polarization. *Business Information Review*, 34(3): 150–160. DOI: 10.1177/0266382117722446.

Sunstein, C. R. (1999) The Law of Group Polarization. *Journal of Political Philosophy*. DOI: 10.2139/ssrn.199668.

Taber, C. S., & Lodge, M. (2006) Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science*, 50(3): 755–769. DOI: 10.1111/j.1540-5907.2006.00214.x.

Tversky, A., & Kahneman, D. (1973) Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2): 207–232. DOI: 10.1016/0010-0285(73)90033-9.

Tversky, A., & Kahneman, D. (1974) Judgment Under Uncertainty: Heuristics and Biases. *Science*. DOI: 10.1126/science.185.4157.1124.

Yeh, W. C., Zhu, W., Huang, C. L., Hsu, T. Y., Liu, Z., & Tan, S. Y. (2022). A new bat and pagerank algorithm for propagation probability in social networks. *Applied Sciences*, 12(14), 6858.