

Report from Dagstuhl Seminar 21402

Digital Disinformation: Taxonomy, Impact, Mitigation, and Regulation

Edited by

Claude Kirchner¹ and Franziska Roesner²

1 CNPEN/CCNE and Inria – Paris, FR, claude.kirchner@inria.fr

2 University of Washington – Seattle, US, franzi@cs.washington.edu

Abstract

We report on the discussions and conclusions of a Dagstuhl seminar focused on digital mis- and disinformation, held in October of 2021. An international and interdisciplinary group of seminar participants considered key technical and societal topics including trustworthiness algorithms (i.e., how to build systems that assess trustworthiness automatically), friction as a technique in platform design (e.g., to slow down people’s consumption of information on social media), the ethics of mis/disinformation interventions, and how to educate users. We detail these discussions and highlight questions for the future.

Seminar October 3–6, 2021 – <http://www.dagstuhl.de/21402>

2012 ACM Subject Classification Information systems → Collaborative and social computing systems and tools; Security and privacy → Human and societal aspects of security and privacy; General and reference → Verification


Keywords and phrases Information, disinformation, misinformation, fake news, deep fake, ethics, trustworthiness, friction, verification

Digital Object Identifier 10.4230/DagRep.11.9.28

1 Executive Summary

Claude Kirchner

Franziska Roesner

License  Creative Commons BY 4.0 International license
© Claude Kirchner and Franziska Roesner

Dagstuhl Seminar #21402 on Digital Disinformation occurred on October 4–6, 2021. The seminar was initially planned by Claude Kirchner (CNPEN/CCNE & Inria), Ninja Marnau (CISPA), and Franziska Roesner (University of Washington), and it was then co-lead and this report was written by Kirchner and Roesner, with input from other seminar participants. The seminar had been originally planned for June of 2020 but was then postponed due to the COVID-19 pandemic. It was held in a hybrid format, with some participants on-site in Dagstuhl and most others joining remotely via the video conferencing system Zoom.

In order to maximize discussion and allow the interests of the group to drive the direction of the seminar, we did not plan for formal talks. Participants were asked to prepare a single slide, few-minute introduction about their research interests and methodologies related to digital disinformation, and a “burning question” they have in the space.



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Digital Disinformation: Taxonomy, Impact, Mitigation, and Regulation, *Dagstuhl Reports*, Vol. 11, Issue 09, pp. 28–44

Editors: Claude Kirchner and Franziska Roesner



DAGSTUHL
REPORTS Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Participants included the following individuals, spanning a range of expertise from computer science to law:

- Esmā Aïmeur (University of Montréal, Canada)
- Jos Baeten (CWI Amsterdam, Netherlands)
- Asia Biega (Max Planck Institute for Security and Privacy, Germany)
- Camille Darche (CNPEN, Inria and Université Paris Nanterre, France)
- Sébastien Gambis (Université du Québec à Montréal, Canada)
- Krishna Gummadi (Max Planck Institute for Software Systems, Germany)
- Claude Kirchner (CNPEN/CCNE and Inria, France)
- Vladimir Kropotov (Trend Micro, Russia)
- Jean-Yves Marion (Lorraine University, France)
- Evangelos Markatos (University of Crete, Greece)
- Fil Menczer (Indiana University, USA)
- Trisha Meyer (Vrije Universiteit Brussel, Belgium)
- Franziska Roesner (University of Washington, USA)
- Kavé Salamatian (University of Savoie, France)
- Juliette Sénéchal (University of Lille, France)
- Dimitrios Serpanos (University of Patras, Greece)
- Serena Villata (CNRS, France)

Based on our preliminary discussions, we identified four topics of interest to many seminar participants: *trustworthiness algorithms* (i.e., how to build systems that assess trust automatically), *friction as a technique in platform design* (e.g., to allow for people to take time and a step back when consuming information on social media), *the ethics of interventions* (e.g., the ethics of blocking or content moderation), and *how to educate users* (e.g., without creating over-skepticism). We then structured the rest of the seminar around four deep-dive conversations on these topics, described in the subsequent sections of this report. Due to the relatively small size of the gathering, and most participants' broad interest in all four topics, we did not break out into smaller discussion groups but rather continued to discuss as a full group.

2 Table of Contents

Executive Summary

<i>Claude Kirchner and Franziska Roesner</i>	28
--	----

Deep Dive 1: Trustworthiness Algorithms

Reflections on Truth versus Trustworthiness	31
Assessing Trustworthiness	31
Technical Approaches	32
Reflections on Harms from Untrustworthy Content	33
Challenging Trustworthiness	34

Deep Dive 2: Friction as a Technique in Platform Design

Argument for Friction	34
Examples and Proposals for Friction in Platform Design	35
Practical and Ethics Considerations	37
Looking Ahead	37

Deep Dive 3: Ethics of Interventions

Which ethics for which purpose?	38
What values are behind moderation tools?	38
How can ethics committees help governing social media platforms?	39

Deep Dive 4: User Education

Role of User Education	40
Characterizing Education Efforts	40

Conclusion and Looking Ahead 41

Participants 44

Remote Participants 44

3 Deep Dive 1: Trustworthiness Algorithms

Our first deep dive conversation focused on the question of how to build systems that assess trustworthiness automatically.

3.1 Reflections on Truth versus Trustworthiness

Helping people to understand the information they have access to in the digital world could be facilitated by algorithms designed to check veracity of statements, to detect incorrect statements, or to find inconsistencies in a given set of knowledge. While automated mis/disinformation algorithms are often framed in terms of such fact-checking, seminar participants repeatedly made the observation that assessing truth can be challenging.

Automating truth assertion is difficult since the level of specification of a statement to check can vary from formal (as in $2 + 2 = 4$) to nearly formal (as in “ $2 + 2 = \text{four}$ ”) to imprecise and complex (such as “COVID-19 is a hoax”). Even for formal statements, we know since Turing and Post that some facts are even undecidable: for some (rather rare) facts there exists no algorithm able to decide if they are true or false. And for complex, informal statements, establishing truth can be hard or impossible, even if we could transform them into formalized statement by explicitly describing all needed details of the context under consideration. The context of particular statement may depend on time, culture, history, education, current availability of verified information or knowledge (e.g., as we have seen scientific knowledge evolving since the beginning of the COVID-19 pandemic), etc.

Moreover, misinformation is rarely something to which we can assign a binary truth value. Often we find a mix of truth with misleading suggestions, implications, attribution, or false context. There is significant nuance that makes it hard even for expert humans to fact-check some claims. Often the best a human fact-checker can do is to observe there there is no evidence in support of a claim, rather than conclude that the claim is false.

Thus, since establishing truth can be very hard (possibly undecidable), participants suggested that we could instead consider automated assessments of trustworthiness. As an example: not everything on Breitbart News or Occupy Democrats is “false”, but these sources have low credibility according to fact-checking organizations.

3.2 Assessing Trustworthiness

Moving, then, from fact-checking or truth assessment to the assessment of trustworthiness, we discussed several aspects:

What is being evaluated? Approaches for assessing trustworthiness might target *content* (e.g., might this image be a deep fake?), *content producers* (e.g., does this website or social media account have a reputation for sharing trustworthy content?), or even *content consumers* (e.g., does this user have a history of re-sharing misinformation?).

Who is assessing trustworthiness? Trustworthiness assessments might be heavily impacted by factors such as the background, education, and biases of the annotators – either humans, or automated processes that may embed such bias in their design.

What makes something trustworthy? Or rather, what signals might a trustworthiness algorithm use to assess content or sources? Participants mentioned potential signals including:

- Adherence to journalistic standards
- Falsifiability (see Popper)
- Accountability, e.g., a history of correcting errors when they are confirmed as such
- Primary knowledge domain or level of expertise of a source being relevant to the topic at hand
- Confidence that an account is not compromised
- User feedback (though this may be manipulated [12])
- Behavior of users who are known to be good at distinguishing mis/disinformation
- Use of techniques (e.g., clickbait, sensationalism) known to play into human cognitive biases

There is a growing number of news and fact-checking organizations that compile credibility metric for sources according to the above criteria, e.g., <https://mediabiasfactcheck.com>, <https://www.newsguardtech.com>, <https://iffy.news>, and <https://factcheck.afp.com>.

Who is the audience for trustworthiness information? We can consider different audiences for the outputs of trustworthiness algorithms. One potential audience is a social media or other platform itself, for use in amplifying or de-amplifying certain content. Beyond this, the targets of such interventions are often end users themselves, e.g., labels on content to help them assess information as they see it. For any user-facing interventions, the designers must consider whether they are targeting users who are receptive to additional information (e.g., conscientious consumers of information trying to make sense of the information ecosystems) or not (e.g., people already convinced of a false conspiracy theory). And in addition to different kinds of people, there are quite different kinds of information. One size will obviously not fit everyone.

Building systems assessing trust automatically will rely (i) on the many factors listed above, but also on (ii) ways to allow people to have time to take a step back (e.g., using friction as described in Section 4) and (iii) on thinking about the many ethics issues concerning the platforms (Section 5).

3.3 Technical Approaches

Participants also brainstormed specific technical approaches that might be used to support trustworthiness assessment or other verification. The observation was made that online platforms are already doing some of this work, but not much about their techniques or algorithms is public or transparent; therefore the research community cannot assess the soundness and appropriate application of platform regulation based on these metrics.

Automated property proving could be used on sufficiently formal statements, possibly with specific pre-processing of them.

The pagerank technique is an early example, initially used by search engines. These types of algorithms could be used on Twitter graph data, for example, starting with initial trustworthiness labels from existing sources.

Cryptographic techniques can be used to trace the provenance of content, identify the sources of content, and/or attest to attributes of content (e.g., metadata like when and where a video was shot, or when and by whom it was published online). While

cryptographic techniques like digital signatures do allow attribution of content to a (possibly trusted) source, participants pointed out that there are limitations to these approaches because we still need a root of trust – e.g., we might be able to verify the provenance of a video, but we still need to know whether we can trust the source.

Deep fake detection is useful, but needs continuous adaptation to technical improvements in deep fake creation (i.e., an “arms race”).

3.4 Reflections on Harms from Untrustworthy Content

As part of this deep dive conversation, participants also reflected on the potential harms of untrustworthy content – both what those harms are, as well as how they might be used to prioritize different types of sources or content for intervention.

Spread and targeting

The current social internet allows both brute force broadcasting (i.e., everyone receives the same (mis)information) as well as highly targeted (mis)information dissemination. Highly targeted information can be sent to many different people, and they will not see the same thing as it will be automatically tailored to their unique profile – a unique phenomenon in our new digital ecosystem. We observed that both mis/disinformation with broad reach, as well as highly targeted mis/disinformation, can create substantial harm, but may require different approaches to combat.

Assessing and prioritizing potential harms

Different types of mis/disinformation may have have different harms. For example, ill-founded distrust in vaccines has clear potential harms [18], while on its face it seems less problematic if some people believe that the earth is flat – though the distinctions are not so simple, as it has been show that beliefs in different conspiracy theories may be correlated. Moreover, different populations may be more vulnerable to certain harms. With a good characterization we could use that to prioritize limited resources. But characterizing and measuring harm in online content could be very hard and remains an active research area (e.g., [31, 21]). Even irrespective of harm prediction, any prioritization that makes a judgment on which harms to which populations are more “important” than others will have ethical impacts.

Long-term impacts of content despite awareness that source is untrustworthy

Even with trustworthiness information, people might not remember the provenance of information they have seen – they might assume something is true even though they were exposed to it from an untrustworthy source. For example, it has been shown in the context of ads [17] that people do not necessarily remember the sources of information. This observation raises a more general related question: how are people influenced by bad information they just see in their information ecosystem but do not engage much with, such as posts they scroll by or ads they think they ignore? Does that information (especially with source provenance later lost) make it into their worldviews? Is this effect worse when people read only catchy headlines, but not the underlying articles?

3.5 Challenging Trustworthiness

We also noted that trustworthiness is challenged in various ways. First, disinformation is fundamentally an attack on the human brain, taking advantage of aspects of human neuro-cognition, culture, behaviors, and potentially fault or non-rational heuristics. Simply surfacing information about trustworthiness to people will not, on its own, suffice to overcome these issues. A second challenge comes from traditional computer security: information maybe altered on purpose, e.g., via the criminal alteration of data stored in compromised computers in order to support disinformation and falsely justify fake news. For example, a participant pointed to recent targeted attacks on research against COVID-19 vaccines [23]. And finally, while AI and ML can provide tools to help combat disinformation (e.g., supporting trustworthiness assessment), their advancement may also increase the volume and the quality of fake news, as well as bringing new disinformation vectors. These challenges compound the complexity of understanding and controlling the context in which we would like to increase trustworthiness, and require for further research.

4 Deep Dive 2: Friction as a Technique in Platform Design

Our second deep dive conversation focused on the topic of “friction” as a technique in (social media or other information) platform design. The idea of friction involves slowing down some human interaction with the platform, either on the information producer side (e.g., limiting down how often someone can post) or on the information consumer side (e.g., limiting how much time someone spends on the platform).

4.1 Argument for Friction

Some participants presented evidence for why the idea of friction might be fruitful. One piece of evidence: while one might imagine that platform behavior might self-regulate, in the sense that higher quality content should become more popular, in fact, the correlation between quality and popularity is very weak as long as people have finite attention (i.e., are unable to see everything that is posted on the platform) [13]. This reality is problematic from the perspective of misinformation, because it makes us vulnerable to bad actors flooding the network with more information. Indeed, there is empirical evidence that adversaries flood the network with lots of volume of problematic content [25, 22].

At the same time, we know from many examples in other domains and in our own lives that what might be good for a person’s well-being in general, and what they want or choose to do in a given moment, are not always aligned, making us vulnerable to addictive online platforms. Related to this, the question was raised: *is there an equivalent to the “privacy paradox” [3] for misinformation?* Seminar participants posited that the answer is likely yes: that people are not as conscientious about consuming information as they say they are or want to be (e.g., they may still click on clickbait content even though they recognize it as clickbait). However, just as the privacy paradox is not truly a paradox (because people are forced into the impossible choice between opting out of crucial online platforms and agreeing to their privacy ultimatums) [24], we posit that in the context of misinformation, people’s actual choices and behavior are deeply influenced by the platforms. In other words, the situation is not so much a paradox as a failure of platforms to design for their users’ well-being.

Towards that end, we thus considered “friction” as a potential tool in platform design to reduce mis/disinformation and increase well-being. In the extreme, as a thought experiment, a platform might have no friction (i.e., anyone can post or view anything) or only friction (i.e., everything is blocked). The latter case is clearly nonsensical, but as a practical matter, so is the former: participants stressed that *platform designs are never neutral*. Particularly because there is finite time and attention that an individual can spend on a platform, the design choices about who is shown what content in what way, and the affordances for how content is posted and shared, naturally shape people’s consumption and production behaviors.

4.2 Examples and Proposals for Friction in Platform Design

We considered existing examples of friction, as well as our own new proposals, for different stakeholders.

Friction for Information Consumers

Consumers are users who use social media or other information platforms to read and/or view content. Examples of friction for consumers include:

- Most generally, the platform’s algorithm for what is shown to whom when is a form of friction for certain types of content.
- Labeling content (e.g., with “false” labels if something was able to be explicitly fact-checked, or with other information, such as about the trustworthiness of the source as discussed in Section 3) as a nudge for people to think about it a second time.
- Minor changes in an interface can have impacts on how people consume information (e.g., the presence or absence of a search box might change people’s ability to act on their preferences in content consumption [16]).
- Helping people limit how much they interact with platforms, e.g., how much time they spend on social media. Existing tools include screen time capping tools (e.g., smartphone apps or browser extensions), including tools with additional incentives (e.g., planting real or virtual trees). A proposal was to surface the climate impact for use of platforms per unit time, as a new type of incentive/motivation.
- *Proposal:* Some kind of platform guidance for helping shape discourse (e.g., like an ongoing project designing a tool to help local actors intervene in mis/disinformation related conversations [26]).
- *Proposal:* Design social media platforms for more intent-driven interactions. Today, people visiting social media sites are inundated with content on all sorts of topics interleaved on their feeds, which can create information overload, reduce motivation to assess information quality [10], and other potential harms. Imagine instead a platform that greets a user with an empty search box, asking them to take an active role in shaping the topics and types of content that they see. However, we note a potential privacy-related side effect here: active interactions with the platform will also provide more information about the user’s interests and habits that could be exploited for targeted advertising or manipulation; thus, any such approach should be coupled with robust privacy protections.
- *Proposal:* Helping users filter the content they want to see, e.g., on Twitter, to engage on academic research topics but not U.S. politics.

Friction for Information Producers

Information producers are those who create new posts containing mis/disinformation or other low-quality or problematic content. Examples of friction for producers include:

- Twitter limits the number of posts per day (though posters may violate this by deleting things they posted earlier). One might also impose other limits, such as rate limits.
- Twitter used captchas to limit posting volume during the 2020 U.S. elections.
- Content moderation in general was considered a general form of friction (e.g., removing posts or account, shadow-banning). The Facebook Oversight Board was mentioned, which evaluates content removal or account banning decisions.
- Other researchers have highlighted the importance of banning “repeat offenders”, accounts that frequently post or spread mis/disinformation [8].
- *Proposal:* On WhatsApp, a message that has spread beyond a certain number of people could be made public, to allow for scrutiny and content moderation (otherwise challenging on an end-to-end encrypted platform).
- *Proposal:* Some kind of monetary cost for posting (similar to what advertisers must do today, or similar to postage stamps for physical messages).
- *Proposal:* Some kind of “friction score” or “trustworthiness score” for information producers, where higher-quality information producers are able to post more, with more reach. The idea would be to create disincentives (cost) for bad actions, and incentives or rewards for responsible use. This proposal comes with many questions and challenges, including how such a score should be computed, by whom, the need to provide transparency about its implementation and enforcement, the privacy risks associated with how the score might be misused, etc. Some ideas for factors that might be incorporated into the score included: the quality of previously shared links, prior track record of posted content that has been flagged or taken down, certain behaviors (e.g., high-frequency posting, inauthentic/coordinated behavior), and audience diversity or other audience properties. Or could we translate some of the trust indicators developed for news organizations (e.g., as used by Newsguard) to individuals? Using crowdsourced data as part of the score was cautioned against, due to risks of coordinated activity or similar [30].
- *Proposal:* Time-based friction for posting, when critical events like elections are happening or shortly before a company goes for IPO, friction could be increased to minimize manipulations at the time when they are most impactful. This idea could apply to friction for information consumers and/or intermediaries.

Friction for Intermediaries

Intermediaries are users who knowingly or unknowingly re-share mis/ disinformation posted by others (e.g., retweeting on Twitter). Examples of friction for intermediaries include:

- WhatsApp limits the number of people to whom someone can forward a message.
- Facebook alerts you if you try to share a link or post that their fact-checkers have labeled as false.
- Twitter warns you if you retweet a link without having first clicked on it.
- *Proposal:* Temporarily disable reposting capability for users who ignore warnings with high frequency.
- *Proposal:* The above-mentioned “trustworthiness score” could be used for intermediaries as well.

- *Proposal:* Behaviorally identifying users who are good at identifying and/or avoiding mis/disinformation, and using that information to, for example, amplify or reduce the spread of certain types of content.

4.3 Practical and Ethics Considerations

Orthogonal to any specific friction proposals, participants voiced concerns about practical considerations in deploying such techniques. First, there are questions about the incentives of the platform designers, in terms of economic models and business incentives. Indeed, after the seminar, there was the example of a browser extension designed to limit people's Facebook use that was shut down by Facebook [20]. The corresponding observation was that these questions are not merely technical challenges but will require regulations to put any solutions into place.

Additionally, there are ethical, legal, and political questions around the power of platforms to make decisions like promoting or demoting certain types of content or blocking certain people. One participant also noted that the use of "friction" is, of course, not ubiquitously a good thing, pointing, for example, to work documenting the uses of friction and flooding as techniques by authoritarian governments to exert control over information in potentially harmful ways [19]. We discuss ethics more in Section 5.

4.4 Looking Ahead

Clearly, significant work remains to be done to design, implement, and evaluate the impacts of different types of friction proposals, compounded with the challenges of platform incentives and their limited transparency and oversight. Stepping back, we concluded our discussion on this topic with a crucial, overarching question: *How might we redesign platforms more radically?* For example, if we wanted to design a "public interest social media platform", what would that look like? Could it be done? Or, looking further into the future towards AR/VR or "metaverse" interactions: can such platforms be designed in ways that help move us more towards (imagined?) ideals of in-person instead of digitally-mediated interactions?

5 Deep Dive 3: Ethics of Interventions

On the second day, we chose to include one formal talk to help seed our discussion on ethics: Serena Villata gave a talk titled "Online disinformation: content moderation, ethical challenges, and future work directions" describing in particular the collective thinking held within the National Pilot Committee for Digital Ethics (Conseil National Pilote d'Ethique du Numérique – CNPEN) [4]. The abstract says:

The purpose [...] is to identify the ethical issues and challenges arising from the widespread use of these different algorithms and tools, which are part of a complex phenomenon with wide-ranging implications. Among others, some questions arise: what does action or inaction in this domain mean in the context of COVID-19? Is it simply a quantitative shift, or are we seeing a more profound change in the nature of the digital solutions designed to fight online disinformation and misinformation?

More generally, how do we face the complexity of this phenomenon, which requires an analysis that seems to go beyond ethics, or even to challenge the notion of ethics itself? Indeed, ethical questions do not arise in the same way depending on whether one is dealing with actors who act consciously to deceive their target or, on the other hand, whether one is dealing with actors who simply get caught up in the flow of information in digital format and, in particular, participate in the virality of this information often in an unconscious way. In the first case, ethical reflection questions responsibility, while in the second case it consists mainly in moving towards awareness. In either case, the requirement is to identify – specifically in the digital domain – the economic, legal, social, political or philosophical dimensions of disinformation or misinformation.

The discussion opened by the presentation helped raise fundamental questions about ethics in the context of mis/disinformation.

5.1 Which ethics for which purpose?

Are there existing ethical frameworks/theories that could be applied in this context?

Shall we develop and/or rely on existing philosophical frameworks? Can we take inspiration from the many AI governance propositions issued from the committees set up, for example, by the UE, IEEE or Unesco? How do we resolve some of the fundamental-seeming tensions (e.g., free speech vs content moderation for safety)?

A crucial point concerns the cultural aspects of information. Do we expect a global set of ethics accepted by everyone, or do we expect ethics to be regional? Clearly, norms and group behaviors are quite often following local customs and they can be influenced or offended by other behaviors in different regions. Related works like the Confucian approaches to tech ethics were mentioned [29].

When looking at ethics as defined by the French Academy [6] (“Reflection on human behaviour and the values on which it is based, carried out with a view to establishing a doctrine, a science of morality.”), the reflection process may evolve over time and of course in different cultures. This highlights the difficulty to answer the question: should we (can we?) establish a common set of rules? And for which purpose?

As ethics precedes regulation and laws that are regional or national in many cases, we see the difficulty, but also the interest, to collaborate on these questions and to try to find or establish common views in the context of global digital platforms.

5.2 What values are behind moderation tools?

When asking which values are behind the moderation tools, the following points were raised.

Algorithmic fairness to avoid issues like algorithmic bias (e.g., the Amazon hiring example).

Bias could in particular be introduced when detecting proactively vs noticing the impacts.

Bias could be better understood and avoided when running simulations to help predict otherwise unforeseen side effects of algorithmic designs.

Transparency appears as a key value here again. For reference, a talk was mentioned from Oana Goga (based on work with Krishna Gummadi and others) auditing explanations provided by social platforms on why some content was shown, highlighting also how these explanations can be manipulated to appear “neutral” [11].

Consent is key as in bioethics, and it relies in part on thoughtful consent design and on education: how to ensure that people understand what they are consenting for? The role of nudging, with much prior work, was also highlighted by participants.

5.3 How can ethics committees help governing social media platforms?

We currently see the setting up of ethical committees close to the platform (e.g., Facebook Oversight Board), possibly acting independently. Can this help? Participants raised the following questions for/about the Facebook Oversight Board:

- What does Facebook do with the user-generated reports of Facebook takeout data (and of course with all the data)?
- What data does the Oversight Board actually get access to?
- Process, data, objectives, consequences (how seriously is this taken)?
- Membership of the board? Multi-stakeholder?
- How does Facebook's internal governance interact with the external committee?
- What about ethics-washing?

Participants also observed:

1. The role of the Oversight Board is limited to content decisions, not actually oversight of Facebook per se, despite its name.
2. Facebook's platform is not neutral: the design of the platform shapes not only consumption behaviors but has also shaped content (e.g., from the Facebook whistleblower interview with 60 Minutes [1], politicians saying they felt forced to take more extreme positions that will work well with Facebook's algorithm)

Every problem Facebook and others are facing is so complex, with conflicting needs from different stakeholders – could we actually do better, if we were there? Are there examples of ethics boards in such companies working well? We left the question open but we begin to think about what would be desirable.

Let's imagine new, different social media platforms

Such a platform might rely on an oversight board accessing information (what?) to make decisions (which?) to be applied. The ethics committee might use tools for being more transparent on data and models without releasing the data and the model, e.g., [9, 15]. And of course it is crucial to involve multiple, globally and otherwise diverse stakeholders.

6 Deep Dive 4: User Education

Finally, we discussed the potential role of and strategies for user education against mis/disinformation. Indeed, many existing educational efforts exist in this space (e.g., several efforts at the University of Washington alone [28, 5], TrendMicro's internet safety educational resources [27], the Fakey game [14]); our goal here is not to suggest that all of the ideas below are new, but rather to summarize the discussion at the seminar about the role of and important considerations in educational interventions.

6.1 Role of User Education

When considering educational interventions, an important question is always to consider how much should the responsibility be on users to protect themselves versus how much should the responsibility be on companies and regulators to obviate the need for education and self-protection. Beyond responsibility, there is a question of how realistic or appropriate it is to expect users to develop deep understandings of the technical systems they use—indeed, many users do not have such understanding [7]. It seems clear to us that the burden cannot and should not be on end users alone. Nevertheless, educational interventions may have an important role to play in a multi-faceted fight against mis/disinformation, and can provide users with agency and empowerment even in the face of slow-moving or otherwise incentivized regulators and platforms.

6.2 Characterizing Education Efforts

Seminar participants considered different ways to characterize or scope educational interventions, and different options for their aims and implementations. These included:

Who is the audience of an education effort? This might include vulnerable populations, such as kids, teenagers, older adults, people with lower literacy. This might include particularly powerful individuals, such as politician, local influencers, C-level executives, business owners, or people in charge of platform design and implementation. And this might include everyone, continuously, who interacts with online information ecosystems.

What is the goal of an education effort? At first blush, the core goal of an anti-misinformation educational effort might seem to be to *help people learn how to identify mis/disinformation*. However, participants were quick to point out that one should avoid teaching only skepticism, which risks cultivating cynicism [2] and undermining trust in information ecosystems altogether. Instead, educational efforts should also include the opposite goal: *helping people identify trustworthy information and sources*.

Other potential educational goals articulated by seminar participants included: *helping people think critically* and *helping people learn situational awareness online*, as well as specifically *helping people identify manipulative techniques being used* in the content that they see (e.g., Cialdini's principles of influence or persuasion). Another goal might be to *help people engage with others productively about misinformation* (e.g., teaching skills for how to interact with someone when they have posted something false).

Participants also observed that educational goals towards *online literacy in general* might have benefits regarding mis/disinformation, in terms of helping people understand how information flows online and how mis/disinformation may spread or be targeted at them. For example: educating people about how online systems and information ecosystems work in general; helping people understand how information about them is collected, processed, and used online; helping people understand what it means when they consent to online services.

The question was also raised about whether misinformation education efforts can take lessons from educational efforts in other domains, e.g., cybersecurity.

How to deliver educational interventions? Education might be (for example): embedded as part of platforms themselves (e.g. for a chatbot platform, by the chatbot itself); incorporated into existing courses across different levels of education; presented in stand-alone workshops; incorporated into existing or stand-alone websites; presented as public service announcements

(e.g., similar to public health messaging by public health agencies); included as part of entertainment media (e.g., TV shows); or delivered through educational games or apps (e.g., [5, 14]).

7 Conclusion and Looking Ahead

This Dagstuhl workshop, occurring right after the main assault of the COVID-19 pandemic, was held in hybrid mode. It was quite different from “standard” Dagstuhl meetings which benefit from more informal and direct in-person interactions. Still, thanks to our engaged participants and the quality of the technical video conferencing support provided by the Dagstuhl organisation, we found our conversations fruitful, with much involvement and interactions between geographically (and time zone) distant participants.

It is clear that there are more questions than answers about mis/disinformation in our online ecosystems at this time, and we add our voices to the growing and interdisciplinary research community in this space. We look forward to future – hopefully fully in-person – Dagstuhl seminars on this topic, as well as future collaborations between seminar participants and within the research community more broadly.

References

- 1 60 Minutes. Facebook Whistleblower Frances Haugen: The 60 Minutes Interview, October 2021. https://www.youtube.com/watch?v=_Lx5VmAdZSI.
- 2 Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty, Elissa M. Redmiles, Meeyoung Cha, and Krishna P. Gummadi. Analyzing biases in perception of truth in news stories and their implications for fact checking. *IEEE Transactions on Computational Social Systems*, 2021.
- 3 Susanne Barth and Menno D.T.de Jong. The privacy paradox – investigating discrepancies between expressed privacy concerns and actual online behavior – a systematic literature review. *Telematics and Informatics*, 34(7):1038–1058, 2017.
- 4 Comité National Pilote d’Éthique du Numérique. Ethical issues in the fight against disinformation and misinformation. Ethics oversight bulletin 2, CNPEN/CCNE, July 2020. https://www.ccne-ethique.fr/sites/default/files/cnpn-dsinf_eng.pdf.
- 5 Chris Coward, Jin Ha Lee, Lindsay Morse, and Travis Windleharth. Misinformation escape room. <https://tascha.uw.edu/projects/misinformation-escape-room/>.
- 6 Dictionnaire de l’Académie française. éthique. <https://www.dictionnaire-academie.fr/article/A9E2876>.
- 7 doteverone. People, power and technology: the 2018 digital understanding report, April 2018. <https://doteveryone.org.uk/2018/04/people-power-and-technology-the-2018-digital-understanding-report/>.
- 8 Election Integrity Partnership Team. Repeat Offenders: Voting Misinformation on Twitter in the 2020 United States Election, October 2020. <https://www.eipartnership.net/rapid-response/repeat-offenders>.
- 9 Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. Datasheets for datasets, 2020.
- 10 Christine Geeng, Savanna Yee, and Franziska Roesner. Fake news on facebook and twitter: Investigating how people (don’t) investigate. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2020.
- 11 Oana Goga. Investigating ad transparency mechanisms in social media, October 2017. <https://lig-membres.imag.fr/gogao/presentations/EJC-Montreal.pdf>.

- 12 Lion Gu, Vladimir Kropotov, and Fyodor Yarochkin. The Fake News Machine: How Propagandists Abuse the Internet and Manipulate the Public. A TrendLabs Research Paper, 2017. https://documents.trendmicro.com/assets/white_papers/wp-fake-news-machine-how-propagandists-abuse-the-internet.pdf.
- 13 Filippo Menczer and Thomas Hills. Information overload helps fake news spread, and social media knows it. *Scientific American*, December 2020. <https://www.scientificamerican.com/article/information-overload-helps-fake-news-spread-and-social-media-knows-it/>.
- 14 Nicholas Micallef, Mihai Avram, Filippo Menczer, and Sameer Patil. Fakey: A game intervention to improve news literacy on social media. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021.
- 15 Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Jan 2019.
- 16 Nuno Mota, Abhijnan Chakraborty, Asia J. Biega, Krishna P. Gummadi, and Hoda Heidari. On the desiderata for online altruism: Nudging for equitable donations. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), oct 2020.
- 17 Michel Tuan Pham and Gita Venkataramani Johar. Contingent Processes of Source Identification. *Journal of Consumer Research*, 24(3):249–265, 12 1997.
- 18 Francesco Pierri, Brea Perry, Matthew R. DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, and John Bryden. The impact of online misinformation on U.S. COVID-19 vaccinations. *CoRR*, abs/2104.10635, April 2021.
- 19 Margaret Earling Roberts. *Fear, Friction, and Flooding: Methods of Online Information Control*. PhD thesis, Harvard University, 2014.
- 20 Lucas Ropek. Facebook Banned the Creator of ‘Unfollow Everything’ and Sent Him a Cease and Desist Letter, October 2021. <https://gizmodo.com/facebook-banned-the-creator-of-unfollow-everything-and-1847826505>.
- 21 Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. A framework of severity for harmful content online. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), October 2021.
- 22 Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 9, 2018.
- 23 Jon Sharman. Hackers targeted University of Oxford’s Covid vaccine research, cyber spies reveal, November 2021. <https://www.independent.co.uk/news/uk/home-news/covid-vaccine-hack-cyber-oxford-b1959147.html>.
- 24 Daniel J. Solove. The myth of the privacy paradox. 89 *George Washington Law Review* 1 (2021), GWU Legal Studies Research Paper No. 2020-10, GWU Law School Public Law Research Paper No. 2020-10, January 2021. <https://ssrn.com/abstract=3536265>.
- 25 Pablo Suárez-Serrato, Margaret E. Roberts, Clayton Davis, and Filippo Menczer. On the influence of social bots in online protests. In Emma Spiro and Yong-Yeol Ahn, editors, *Social Informatics*, pages 269–278. Springer International Publishing, 2016.
- 26 Nevin Thompson. Hacks/Hackers, Partners Awarded Funding to Participate in the 2021 National Science Foundation’s Convergence Accelerator, September 2021. <https://newsq.net/2021/09/22>.
- 27 TrendMicro. Internet safety for kids & families. <https://www.trendmicro.com/internet-safety/>.
- 28 University of Washington Center for an Informed Public. Misinfoday. <https://www.cip.uw.edu/get-involved/misinfoday/>.

- 29 Pak-Hang Wong. google scholar page. https://scholar.google.com/citations?hl=en&user=e4yJCwcAAAAJ&view_op=list_works&sortby=pubdate.
- 30 Taha Yasseri and Filippo Menczer. Can the wikipedia moderation model rescue the social marketplace of ideas? *CoRR*, abs/2104.13754, 2021.
- 31 Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. What makes a “bad” ad? user perceptions of problematic online advertising. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2021.

Participants

- Camille Darche
French Nat. Pilot Committee for
Dig. Ethics – Paris, FR
- Lynda Hardman
CWI – Amsterdam, NL
- Jean-Yves Marion
CNRS – Nancy, FR
- Claude Kirchner
INRIA – Le Chesnay, FR

Remote Participants

- Esmā Aimeur
University of Montreal, CA
- Jos Baeten
CWI – Amsterdam, NL
- Asia J. Biega
MPI-SP – Bochum, DE
- Sébastien Gambs
University of Montreal, CA
- Krishna Gummadi
MPI-SWS – Saarbrücken, DE
- Vladimir Kropotov
Trend Micro – Garching, DE
- Evangelos Markatos
FORTH – Heraklion, GR
- Filippo Menczer
Indiana University –
Bloomington, US
- Trisha Meyer
Free University of Brussels, BE
- Franziska Roesner
University of Washington –
Seattle, US
- Kavé Salamatian
University of Savoie – Annecy le
Vieux, FR
- Juliette Sénéchal
University of Lille, FR
- Dimitrios Serpanos
ATHENA Research Center –
Patras, GR
- Serena Villata
Université Côte d’Azur –
Sophia Antipolis, FR