JeDEM 14(2): 109-129, 2022
ISSN 2075-9517
http://www.jedem.org
Date of submission: 30.06.2022
Date of acceptance: 10.11.2022
Article DOI: 10.29379/jedem.v14i2.746

Towards a Positioning Model for Evaluating the Use and Design of Anti-Disinformation Tools

Mattias Svahn[1*], Serena Coppolino Perfumi[2]

[1] ORCID Nr: 0000-0002-1317-2296*
Department of Computer and Systems Science, Stockholm University, svahn@dsv.su.se

[2] ORCID Nr: 0000-0003-1481-2918
Department of Sociology, Stockholm University, serena.perfumi@sociology.su.se

Abstract: With the increasing amounts of mis- and disinformation circulating online, the demand for tools to combat and contain the phenomenon has also increased. The multifaceted nature of the phenomenon requires a set of tools that can respond effectively, and can deal with the different ways in which disinformation can present itself, In this paper, after consulting independent fact-checkers to create a list, we map the landscape of tools available to combat different typologies of mis and disinformation on the basis of three levels of analysis: the employment of policy-regulated strategies, the use of co-creation, and the preference for manual or automated processes of detection. We then create a model in which we position the different tools across three axes of analysis and show how the tools distribute across different market positions.

Keywords: Disinformation, anti-disinformation tools, fact checking, co-creation, policy

Acknowledgement: This work was initiated within the frames of the Horizon 2020 project Co-inform

1. Introduction and problem discussion

The issue of misinformation in social media is currently attracting a lot of attention, especially for the effects that it has on health-related and political behaviour, among other realms of interest [1, 4]. The Covid-19 pandemic, the war in Ukraine, and the cycle of elections in open democratic societies makes it even more urgent, as the struggle for the narrative becomes real. Misinformation and disinformation are not new phenomena, as shown, for example, by the work of Burkhardt [8], who presents a history of fake news, from Babylonian times through Roman times up until 2017, the dawn of the current disinformation era.

In the same way, also research on misinformation dates back to the mid 1900s, with studies like the one conducted by Allport and Postman [2] on the "basic law of rumour", demonstrating that the strength of a rumour is dependent on the importance of the subject and individual concerns regarding this, as well as of the time and ambiguity of the evidence on the topic.

While the phenomenon of disinformation has deep roots, the social media era has made it more pervasive. The 2000s have witnessed a rapid development of social media platforms that have facilitated the spread of both information and misinformation regarding everything from local to global issues. Studies analysing misinformation on social media platforms have found that misinformation and disinformation travel faster than trustworthy information [4]. This can happen for several reasons. In the age of many-to-many information, each user can create content and, through social media, reach a large number of readers [38], hence each individual's agency in content creation has increased significantly, changing the balance with the type of information spread by traditional sources of information, such as newspapers. Consequently, the amount of information available to each digital media user has increased drastically [36, 37], and therefore, the thoroughness in analysing each piece of information has decreased. Secondly, disinformation in itself has evolved, in the sense that on one side it tends to mimic the language and structure of online newspaper articles, and on the other side it has also started using emotional appeals to catch the attention of the users and stand out [38].

Lately, we have also witnessed the appearance of new forms of disinformation, which do not employ only written articles. One example of this typology of content is deepfakes, which are highly realistic videos that can manipulate the movements and voices of actors present in them to make them resemble realistic videos [40]. Examples of deepfakes have involved highly influential figures such as Barack Obama, Donald Trump, as well as journalists delivering news or tech moguls announcing technological innovations. Deepfakes are incredibly realistic, and very hard to spot for the users, but also for neural networks and the other available technologies in place to assess the trustworthiness of the content of the videos [28]. For these reasons and for the threat that deepfakes pose, research on the topic has been trying to understand the phenomenon and how to counter it. For example, Rana and colleagues [40] conducted a review of the current deepfake research, and Deshmukh and Wankhade [11] conducted a review of current deepfake detection technologies.

Recent research on misinformation and disinformation has focused on many aspects of the phenomenon. Di Domenico and colleagues [12] conducted a review, and through critical evaluation and synthesis of the literature, identified five themes that explain the fake news phenomenon: the dissemination process, spreading channel features, outcomes, fabricated legitimacy, and attitudes.

Other studies have focused on taxonomizing the phenomenon for example Choy [9] proposed a framework for identifying "fake news". Another example is the work of Giglietto and colleagues [19], which is based on factors, such as perceptions of the source, the story, the context, and the decisions of the audience and the propagator. The authors propose a taxonomy of "pure disinformation" where both the original author and the propagator are aware of the "false" nature of information, but they nevertheless decide to share it. A different situation is represented by the typology called "misinformation propagated through disinformation" where information is originally produced as "true" and then shared by a propagator who believes it is "false". Finally, "disinformation propagated through misinformation" is the situation in which information is devised as "false" by a creator but is perceived as "true" by a propagator. Research on misinformation in the media has been progressing rapidly as well, with studies contextualising misinformation within different realms and events, such as within journalism [13], in the context of elections [50], and also trying to predict and identify future challenges [16].

The pace and evolution of mis- and disinformation had social media platforms and their architecture at the centre of discussion on the spread of misinformation. The echo-chamber-like structure of the networks within social media platforms [17, 36], and the personalization process carried out by algorithms, can reinforce existing biases within the users [37], who tend to be exposed mostly to information that supports their pre-existing beliefs. Thus, platforms like Facebook, Twitter, and Instagram have started collaborating with fact-checkers to flag incorrect information and to give users the possibility to report items for fact-checking. However, the tools provided by these platforms do not yet respond effectively to the demands [17].

To counter the spread of misinformation and disinformation, fact-checking work, namely the act of taking up published information, examining it for factualness and veracity, and re-publishing it, has also been carried out intensively in the past years. Many fact-checkers are members of the International Fact Checkers Network, -IFCN[1], which provides assistance to the biggest social media platforms in the evaluation of the posted content. Furthermore, given the demand for fact-checking, several tools have been developed to help users navigate the information landscape within social media. These tools deal with different typologies of false information, and use different strategies to detect, flag, and select the items, but there is no model to evaluate what is available, how the tools function, what is missing from the market, and how the tools could evolve. For these reasons, with this work, we aim to build a model that can be used to understand these aspects in relation to the available anti-misinformation tools, as well as to provide indications on where the market is moving and what is missing from the current landscape.

The aim of the proposed model in this paper is to map the current anti-disinformation tools landscape, by analysing the architectural choices that govern the tools' functioning and response to different types of disinformation. In this way, we aim at observing the features of the most widespread tools on the market in order to observe the common qualities of tools dealing with different forms of misinformation and disinformation, as well as potential future directions in the development of anti-disinformation software.

We see tools as software developed with the intention to detect and in some way judge and give the user a notice of mis- or disinformation. In this sense, the tools as intended in this work, can take the form of plug-ins, browser extensions, downloadable software, notifications of flagging appearing on social media platforms, and website-based software. Repositories of fact-checked information have not been taken into account as they are often the source these tools operate on, and cannot be categorised as software.

[1] https://www.poynter.org/ifcn [Accessed 2022-11-03]

It is also noteworthy, that the recent Oxford Reuters Institute's trend report [35] pointed out that news media organisations around the world are focusing on content in the form of podcasts, newsletters, and videos, but are not working on developing anti-disinformation platforms integrated within the news media infrastructure, hence for the present and the near future, tools for fact-checking will often be separated from the news media platforms as the ones presented in this study are.

In this way, we provide a framework for understanding what is available to the users, how these tools operate to detect and notify misinformation and disinformation, what is most commonly developed in order to tackle different types of misinformation and disinformation, what is missing within the current landscape, and in which direction the tools should evolve in order to provide diversified options catering to the users' needs, as well as to effectively address different typologies of misinformation and disinformation in different online environments.

2. Literature review on the qualities of anti-disinformation tools

The work on combating disinformation can take on many shapes, at least as many as disinformation itself. Farrell and colleagues [14] outlined some of the specific problems that misinformation detection has to address. The authors also outline some different models of how misinformation spreads that are relevant for detection, and provide a typology of anti-misinformation tools, such as style-based, knowledge-based, propagation-based, or credibility-based tools. They position tools within the misinformation ecosystem, with regard to how, when, and what kinds of misinformation they handle. However, Farrell and colleagues do not focus on the users' perspective, a gap that is taken up by the work of Komendatova and colleagues [27]. The authors reviewed disinformation tools, focusing on design approaches, by putting them into a perspective of value-driven design. They found that design qualities of a lean-back character, namely not favouring active engagement, are preferred by stakeholders if compared to approaches favouring user engagement.

In the discourse on the qualities of anti-disinformation tools, it is common to refer to artificial intelligence (hereafter AI, and/or machine learning (hereafter ML). Sometimes the participants in the discourse do not always pause to define just what they mean with these terms. We in this study will also use these two terms liberally. AI is a term that has been intensely discussed for a long time. We subscribe to the definition by Samoili and colleagues [46] who see AI as:

> "*Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans(2) that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.*"

We find that definition useful as a description of the AI quality in the anti-disinformation tools that is the object of research for this study. We see ML as a pathway to AI. ML uses algorithms to automatically learn insights and recognize patterns from data, applying that learning, so an AI system will make increasingly better decisions [11]

Many studies have focused on the use of AI to counter disinformation in online spaces. AI has been identified as a powerful and cost-efficient tool to identify online disinformation, as it can quickly analyse a vast amount of items and provide outputs on their nature [6].

AI solutions against disinformation can identify and remove content of a different nature, but it has also been used successfully to detect fake accounts at the origin of disinformation as the account cascades on social media platforms. Furthermore, AI-based tools have also been able to identify linguistic patterns present in already flagged items, in order to facilitate the identification of other disinformation content [26].

There are, however, limitations related to the use of AI. One of the most common ones is related to the precision of AI technologies and the risk of flagging or removing content that is accurate (i.e., false positives). This issue presents concerns when it comes to censorship and freedom of expression as expressed by e.g. Gunton [21]. The second main issue is also related to inaccuracies in AI technologies, as they often are built to include human-like biases, which can return outcomes that over-target specific social groups or present them negatively. These issues are also hard to detect, as tracking down the specific functioning of algorithms often goes beyond the knowledge of the people who develop them [26].

At the time of writing AI technologies are considered efficient in detecting disinformation, but they are still mostly used in combination with manual solutions, to try to limit the errors that might arise if, fully-automated only, procedures were embedded and implemented in online spaces.

This multitude of shapes and qualities that anti-disinformation tools can assume, impacts the ontology of the theoretical model we propose. Still, as some qualities seem to be present across many of the tools, and refer to how they are constructed, how they function, and what they allow means an ontology can still be made. Within the main qualities that have been identified, the tools we examined in this study can rely on either automated or manual fact-checking, which means that the veracity of the information is assessed either by algorithms, that rely on the existing databases of fact-checked information, or by humans who check the claims against other verified sources of information. They can be either proactive or reactive (i.e., what we will call 'policy' versus 'ad hoc'), meaning that they either set up strategies to prevent the circulation of misinformation, ex-ante or they react to misinformation at the point when it is detected in the system, ex-post. Or finally, the tools can, to varying degrees, rely on collaborative, co-creative efforts. This means that some tools allow the broader community of users to provide information in different forms, such as flagging, fact-checking or feedback on the fact-checking process.

Our object of research is the ontology of design qualities, that make up the conceptualization space of anti-disinformation tools. We do that by organising a theoretical model of the conceptualization space. With this, we take a perspective of critical realism and constructionism [33].

2.1. Aspects of anti-disinformation tools

As highlighted in the previous paragraph, three main qualities of the existing anti-disinformation tools indicate how these approach disinformation (proactively or reactively, namely establishing policies a priori, or establishing ad-hoc strategies to respond to specific cases), which modality they use to analyse it (manual or automated) and to what extent they allow joint efforts (degree of co-creation). As highlighted in the previous paragraph, three main qualities of the existing anti-disinformation tools indicate how these approach disinformation.

It can be done proactively or reactively, namely establishing policies a priori, or establishing ad-hoc policy in order to respond to specific cases. It can be which work modality (manual "handicraft" or automated – AI/ML) the tools employ to analyse disinformation. It can be to what extent the tools allow joint efforts between tool and user (degree of co-creation).

We have, in establishing this found Babakar [3] inspiring, as they postulate that fact-checking exists in a triangular trade-off where the angles are, first; Speed: how quickly the task can be done; second; Complexity, and third; Difficulty: how difficult the task is to perform. Babakars´ model can be seen as a precursor to the model this study presents. Babakars´ triangle corner of Speed versus Difficulty relates to this study's axis of AI versus handicraft and their third angle of Complexity can relate to this study's axis of policy versus ad hoc evaluations. The value of Babakars´ study is the way it, regardless of triangles or axes, launches a notion of how a systems designer of anti-disinformation tools can only optimise for a limited number of qualities at a time. That informed our notion that an analysis of the trade-offs made by designers of today's tools can illustrate where the near future development of anti-disinformation tools needs to move toward. Therefore our studies´ cube model postulates that inherently opposed trade-offs exist, on the basis of all these identified qualities, illustrated. Hence, we proceed to build our model on the basis of three axes of analysis: degree of policy, degree of co-creation, and degree of AI. These three axes can be seen as three semantic differential scales. We identify these categories as crucial in understanding the functioning and efficacy of anti-disinformation tools, as they relate to different aspects of their functionality: speed in detecting, the possibility to outsource efforts, the possibility to engage the users and allowing them to provide opinions and discussion, and which typologies of disinformation and misinformation they focus on. In this way, we aim at mapping the market to observe which of these qualities are more common, which are not, and in which direction the market could, and should go, to present a comprehensive asset of tools for different needs and contexts. Below we present a description of the three degrees that constitute the axes of our model.

2.1.1. Degree of Policy

The X-axis of our model is dedicated to assessing to what extent a tool's actions are driven by a pre-set policy. The degree of policy is measured in high and low, where policy represents the high end of the spectrum, and ad-hoc represents the opposite, low end. Research in the area of policy for anti-disinformation tools has shown that policy work is crucial in a user-focused conceptualization of such tools [27]. It is crucial for the value and structure of policy when regulating anti-disinformation work, and for the value of structured information when devising a user-focused conceptualization [43]. Han and colleagues [22] proposed a git-based framework for developing platform policies on misinformation in a decentralised and collaborative way and suggested a methodology for testing policy settings for anti-disinformation tools. The many and varied taxonomies of mis- and disinformation creates a complex landscape and contribute to the tapestry of definitions for the two phenomena [24]. Wardle [50] identified six different types of misinformation [51] which laid the foundation for the mis/disinformation taxonomy, furthered by the work of Farrell and colleagues [16], and Burgoon and colleagues [7] who discussed misinformation in terms of deceptive language and false context. It is important to note this multitude of taxonomies, as it contributes to showing that awareness of policies for determining mis- and disinformation is both necessary and advantageous when devising a user-centred approach to anti-disinformation tools. With the axis, degree of policy' in this study we measure, to what extent the various anti-disinformation tools on one hand, establish rules that regulate the definition and circulation of misinformation ex-ante, and an instance of misinformation detected according to these policies triggers the tool, which provides a solution to it, or different solutions, according to the pre-established policies. On the other hand, there are tools that evaluate circumstantially the single cases, therefore, reacting to misinformation or disinformation only when encountered. Hence a high-end place on the spectrum represents a tool design based on an ex-ante policy that drives the judgments of the tool, a low end is a situation of a more flexible character that evaluates instances of misinformation sui generis, post facto.

2.1.2. Degree of Co-creation.

The Y-axis of our model is dedicated to assessing to what extent a tool allows for, and, in the extreme end is even driven by, co-creation. This is because in the last few years co-creation has spread rapidly in the business sector, as a way of engaging with stakeholders and building knowledge [1]. The application of co-creation methods is more recent in the public sector, particularly for policy development, and multiple challenges still need to be overcome [29]. It has been suggested that the co-creation of anti-misinformation work in the public sector can be a way of meeting the multifaceted complexity of the task of anti-misinformation work [33], and Osborne and colleagues [35] argue for co-creation to be seen as one element in a larger model of value creation for the public sector. Komendantova and colleagues [27] argue the importance of co-creating the design of anti-disinformation tools together with stakeholders. Co-creative efforts can be particularly valuable in the fight against disinformation, as this becomes a joint effort from fact-checkers, policymakers, developers, and private users, who can all take part in the detection and correction of misinformation and disinformation at different stages.

In our model, co-creation shall be seen as the extent to which the tools allow contributions and inputs from larger communities of users while carrying out the fact-checking work. This can, for example, take the form of allowing the users to flag or report information they encounter on social media to be fact-checked because it looks suspicious, to provide feedback on incorrect labelling of information (e.g., signalling false positives or false negatives), or providing feedback on the functioning of the tool.

In our model we therefore, measure the level of co-creation, where a high degree of co-creation represents the high end of our second axis, where the user of a disinformation tool can, in various ways, contribute to the tools evaluations, and judgments, and low degree of co-creation (solo) represents the opposite, low end of the same spectrum, where a tool is more of a “black box” not allowing for any input from the user.

2.1.3. Degree of AI/ML

The Z axis of our model is dedicated to the assessment of the extent to which the misinformation detection process is carried out, by means of manual or automated fact-checking, applying AI and ML as defined by Samoili and colleagues [46] earlier in this article. The increasing amount of information that requires fact-checking produces an increasing demand for a number of tools and fact-checking services, that in order to cope with sheer volume, rely wholly or partially on the use of AI and machine learning to assess the veracity of information.

Schuster and colleagues [42] built a system that teaches an algorithm to check for spelling errors and factual errors. That system is contrastive, namely, it relies on evidence pairs that are nearly identical in language and content, with the exception that one supports a given claim while the other does not this seems to work, and it seems to be successful. Nakov and colleagues [33] find that automated verification of claims seems like the ultimate application of AI to fact-checking. If such technologies can be developed and deployed, they would allow fact-checking organisations to be faster and to provide more comprehensive coverage than manual fact-checking could ever achieve. However, Nakov and colleagues [33] find, in dialogue with real fact-checkers, that fact-checking is a complex and nuanced process. Claims are not simply correct or incorrect, but may be partially correct, or technically correct but misleading in the wrong context, etc. Hence fact-checking is inherently done, with as much nuanced contextual evaluation as human communication in general and that is not a perfect fit for current AI and ML. Still, given the fact-checking demands, initiatives to combine the two modalities are being considered, and the British Fact-Checker Full Fact has an in-house initiative to build an automated fact-checking system [18].

In today's tools, AI/ML and human evaluation, therefore, exist side by side on a scale. It is an approach that can either be used as the sole method of information scanning, or in conjunction with manual fact-checking, either in an in-house capacity at the fact checker or outsourced to co-creation, to increase capacity. Given the discussions on the matter and the variety of tools that use one approach, the other, or a mixed approach, in our model we measure the degree of AI, where AI/ML represents the high end of the spectrum, characterized by sole use of AI and ML, and manual represents the opposite, low end of the spectrum, where the fact-checking work is carried out only manually.

3. A model for evaluating anti-disinformation tools for conceptual purposes

The presented qualities can be imagined as axes and presented as a model in the form of a cube. This form of a model also illustrates how we postulate the three qualities to be communicating buckets, where a movement on one of the scales, means a movement also on the other two. We apply the examplesbelow as a means of validation of the axes in the cube.

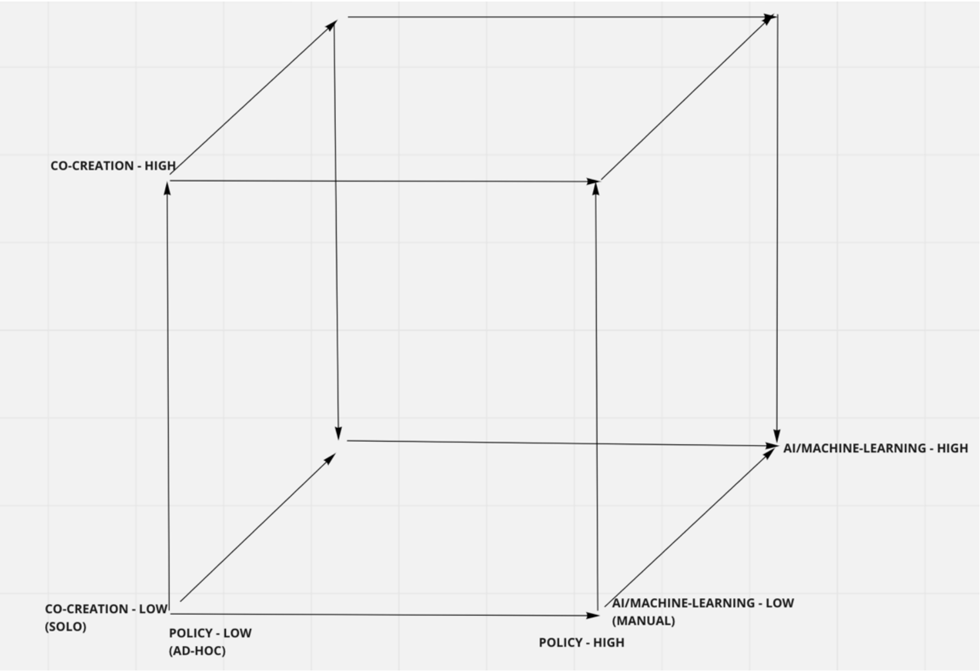

Figure 1 Organisation of the axes in the cube.

As shown in Figure 1, the three axes can visually constitute a three-dimensional cube, which for the building of our model, will then be populated with the selected tools, in order to identify their positioning across the different dimensions of interest. In this way, it will be possible to visualise how the different tools score in terms of employment of co-creation versus a stand-alone design or something in-between; how they regulate their functioning, namely if they have a clear policy or if they vary according to the specific situation, and finally if they employ mostly AI-solutions if most of the process is carried out manually, or a combination of the two. Together the three-dimensional placement of a tool and its place versus other tools, and the groupings of other tools will tell a reader what the current map of anti-disinformation tools is, what design elements are underserved by the current tools on the market, and what design elements are over served. This principle is inspired by the min-max equilibrium theory product differentiation of Ansari and colleagues [3].

4. Review of current anti-disinformation tools

As mentioned previously, in the past few years several tools that rely on both automated and manual fact-checking, or on a combination of the two modalities have been developed to help users navigate the information encountered online. The overarching aim of these tools is to fact-check different aspects of the disinformation content, such as the information architecture (e.g. the claims, the sources, the authors, and the platforms), and aspects related to the typology of content (e.g., images, headlines). These tools can be employed as a way to prevent the spread of disinformation, or as a way for a user to react to it. Here follows a short review of these before we move into populating the cube.

Among the tools that more generally assess the content are; NewsGuard[2] also analysed by Mensio and colleagues [31], Claimbuster[3], The Factual[4], CredEye[5], Public Editor[6], Newstrition[7], and CrowdTangle[8]. Within these, tools like Cyabra[9] employ more advanced technologies to detect deep-fakes, and the Co-Inform Dashboard[10] is designed to help professionals such as journalists, fact-checkers, and policy-makers in their everyday job.

Some social media platforms have developed their own fact-checking systems, often in collaboration with external fact-checking organisations. Examples of these are the Facebook Fact-Checking Program[11], Twitter Birdwatch[12], and the Whatsapp IFCN chat bot[13].

Besides the tools developed in collaboration with the platforms, there are also platform-specific tools developed externally, like CaptainFact[14], designed for YouTube, Foller.me[15], the Co-Inform plug-in[16] and Hoaxy[17], designed for Twitter, FakeSpot[18], specific for e-commerce platforms, and tools like Botometer[19] that are not only platform-specific but also agent-specific: Botometer provides information about the probability that a Twitter account is a bot. Other tools are content-specific, for example, TinEye[20] and RevEye[21] focus on detecting fake and decontextualized images, the WeVerify/InVID[22] plug-in is created for the assessment of videos, and Sensity.ai[23] identifies fake users and frauds.

Finally, tools are being developed not only to detect, signal, and correct misinformation, but also to work on the critical thinking and analytical abilities of the users, and one of these is Fiskkit[24]. To

[2] https://www.newsguardtech.com/ [Accessed 2022-10-17]

[3] https://idir.uta.edu/claimbuster / [Accessed 2022-10-17]

[4] https://www.thefactual.com/ [[Accessed 2022-10-17]

[5] https://dl.acm.org/doi/fullHtml/10.1145/3184558.3186967 [Accessed 2022-10-17]

[6] https://www.publiceditor.io / [Accessed 2022-10-17]

[7] http://newstrition.github.io/newstrition/ [Accessed 2022-10-17]

[8] https://www.crowdtangle.com/ [Accessed 2022-10-17]

[9] https://cyabra.com/ [Accessed 2022-10-17]

[10] https://coinform.eu [Accessed 2022-10-17]

[11] https://www.facebook.com/journalismproject/programs/third-party-fact-checking [Accessed 2022-10-17]

[12] https://twitter.github.io/birdwatch/about/overview / [Accessed 2022-10-17]

[13] https://www.poynter.org/fact-checking/2020/poynters-international-fact-checking-network-launches-whatsapp-chatbot-to-fight-covid-19-misinformation-leveraging-database-of-more-than-4000-hoaxes/ [Accessed 2022-10-17]

[14] https://captainfact.io [Accessed 2022-10-17]

[15] https://foller.me/ [Accessed 2022-10-17]

[16] www.coinform.eu [Accessed 2022-10-17]

[17] https://hoaxy.osome.iu.edu/ [Accessed 2022-10-17]

[18] https://www.fakespot.com/ [Accessed 2022-10-17

[19] https://botometer.osome.iu.edu / [Accessed 2022-10-17]

[20] https://tineye.com/ [Accessed 2022-10-17

[21] https://chrome.google.com/webstore/detail/reveye-reverse-image-sear/keaaclcjhehbbapnphnmpi-klalfhelgf [Accessed 2022-10-17]

[22] https://www.invid-project.eu/tools-and-services/invid-verification-plugin/[Accessed 2022-10-17]

[23] https://sensity.ai [Accessed 2022-10-17]

[24] https://fiskkit.com/ [Accessed 2022-10-17]

make sure that fact-checking work is carried out transparently and to counter the rise of counter-fact-checking initiatives, the International Fact-Checking Network provides information and assessments on the credibility of the various fact-checkers and fact-checking initiatives.

5. A model for evaluating the concept space of anti-disinformation tools.

We have proposed a model to be used to categorise, define and evaluate anti-misinformation tools from the perspective of the functionalities encountered by the user. We have also reviewed a number of current anti-disinformation tools used by fact-checkers. Here we map a number of chosen examples to the cubic model.

In late 2020 we created an initial theoretical sample and made a theoretical design analysis. In February 2021 we tested the content validity of the choice and initial analysis against a panel of nine experts from the IFCN network; those results were presented at the IFIP EGOV2021conference[25] in 2021 [46].

In February 2022, on the basis of the outcomes of the initial testing, we developed a new series of structured questions coming out of the work from 2021. These questions were meant to further explore constructs relating to the user experience of the design of the tools and the axes and ask for examples and explore new market entrants that may perhaps not have been covered the first time. These new questions were presented to the experts involved in the first testing, and then to four more experts from the IFCN network in order to validate the model, bringing the number of experts interviewed for the purpose of this study to thirteen. In the structured interviews, the respondents were, among other things, asked about how they evaluated the tools' qualities on the three axes. The 2022 result then complements the theoretical design analysis and empirical work initially made in 2020-2021.

The second round of interviews with IFCN experts yielded seven new examples of tools. Out of these seven examples four were discarded. The three remaining were plotted onto the cube with some following adjustments of the original model from 2020-2021. The three newly selected tools from 2022, that were not included the first time, are CrowdTangle, RevEye, and Sensity.ai. These are included as all three match our definition of tool. Four of the newly presented examples were excluded from our analysis, as they did not match our definition of tool. Two of the four excluded were Trendolizer[26] and Google Fact Check[27]. These were excluded as they are repositories of disinformation rather than tools. The two others are Yandex[28] and Advanced Google Search[29] which are search engines that are not built specifically for anti-disinformation and anti-misinformation purposes, and therefore, fall outside the definition for this study.

[25] https://dgsociety.org/egov-2021/[Accessed 2022-10-17]

[26] http://www.trendolizer.com/ [Accessed 2022-10-17]

[27] https://toolbox.google.com/factcheck/explorer [Accessed 2022-10-17]

[28] https://yandex.com/[Accessed 2022-10-17]

[29] https://www.google.com/advanced_search [Accessed 2022-10-17]

The chosen tools were placed within the cube in a place that represents their position in relation to the three axes, as mapped by the initial design analysis in 2021 and the surveys in 2021 and 2022. The bottom horizontal X-axis represents the degree of policy, moving from left (lowest degree) to right (highest degree). The vertical left Y-axis represents the degree of co-creation, moving from the bottom (lowest degree) upwards (highest degree). Finally, the third and last dimension, the Y-axis indicates the degree of AI/ML, and is placed at the conjunction point with the policy axis and is visualised as the cube depth axis, which moves, starting from the conjunction point with the policy axis (lowest degree), and runs alongside the bottom of the right cube face until the next conjunction point (highest degree).

The cubic structure of the model, based on the three aspects or dimensions discussed above and shown in Fig. 2, makes it salient how the axes are connected vessels. To be evaluated, a disinformation tool is measured with a design analysis and positioned in the cube, with respect to the three dimensions, and then evaluated from its position. In the following, final section of the chapter, we apply the proposed model to a number of examples, in order to show whether it can be used as intended.

6. Conclusions and discussion of the model

As we have seen from the literature on misinformation and the anti-misinformation tools, scholars have worked in order to categorise false information, and have studied extensively which factors affect the propagation and perceived credibility of false information. Fact-checking remains, at the moment, the main tool that can be employed to counter the spreading of misinformation and disinformation online. Several fact-checking initiatives have been born in the last years, in order to deal with the different typologies of false information that spread online, as well as with the specific contexts in which false information spreads. Events like the war in Ukraine, the Covid-19 pandemic, and the regular cycle of elections in democratic countries with open societies lead to ever-increasing volumes of mis-and disinformation and increased the urgency to find solutions to the issue of the circulation of false information.

On the basis of these premises, we have created a model that can be used to map the anti-disinformation tools available on the market. The model does so by allowing whoever employs it to categorise the tools according to specific qualities. The model presented above, whilst being theoretical, presents several practical implications for the development and procurement of tools.

We can see a crowded market position in the fifth corner with Hoaxy, Factmata Foller.me, Botometer, Newstrition, and the new tool RevEye. These are all characterised by high degrees of AI/ML, and low degrees of policy and co-creation, which therefore, characterises a form of development that is currently well catered for.

The second most-crowded market position we find in the middle, between the fifth and the sixth corners, it is represented by tools like The Factual, FakerFact, Tin-Eye, and Cyabra. These tools score similarly across our axes, and they are characterised by a high degree of AI/ML, a medium degree of policy, meaning that they use a mixed approach and a low degree of co-creation. A low degree of co-creation is a common denominator for all of these positions and it represents an expression of the current state of the market, despite discourse seeing co-creation as a necessary element [28, 30]. We can expect that, because of the importance given to co-creation by research, its employment might increase in the future.

We can observe that the third corner is empty. It is defined by a high degree of co-creation, a low degree of policy, and a mostly manual process. However, it would be unlikely to see an actual anti-disinformation tool in that market position, as the functioning and usability of such a tool would be impaired by these very characteristics. To provide an example, a platform that perhaps represents this position is Reddit, but it is not within the realm of anti-disinformation tools. The case closest to this position is Twitter Birdwatch as it is co-creative by design but at the time of writing low on policy.

The other tools appear to be more diversified across the various axes, for example, the two other new tools, CrowdTangle and Sensity.ai position themselves in scattered previously empty positions: they both score high in AI/ML, and medium in co-creation, but Sensity.ai scores medium in policy, while CrowdTangle scores high in policy. Some other exceptions are represented by WeVerify (In-VID), the Co-inform plug-in, and the Co-inform dashboard, which are roughly grouped in the same area around the second corner: the Co-inform plug-in and WeVerify are both characterised by high degrees in all the axes (AI/ML, policy and co-creation), while the Co-Inform dashboard also has high degrees of policy and co-creation, but medium degrees of AI/ML. These are among the few that meet up with the idea of applying co-creation to meet the demands of combating misinformation.

It is unlikely that the tools in the crowded fifth corner, defined by being high in AI/ML, low in policy, and low in co-creation will develop away from that corner. This is due to them being designed as stand-alone tools that are not part of any ecosystem. Standing alone is such an intrinsic design quality in these that it is unlikely these will change in the short term. Development along these lines will come from new market entrants.

Birdwatch is an interesting case, it occupies a unique position with being high in human processing and co-creation while being highly dependent on co-created policy. Its future position is dependent on the way it manages to keep and implement co-created policy over time. We can right now in the autumn of 2022 see that the design qualities of Birdwatch may be moving its usage towards the empty corner of a high degree of co-creation, a low degree of centrally coordinated policy, and a mostly manual process. At the time of writing Birdwatch is being expanded from hitherto a prototype to an almost public product and it may perhaps become the first successful entrant in the empty corner.

Generally speaking, we can observe that in recent times there has been an increase in the use of AI and ML in the detection of misinformation and disinformation, and this is due both to the increased amount of false information circulating on social media platforms, but also to the nature of it. AI and ML can indeed facilitate the detection process by analysing vast amounts of information in a short amount of time. Furthermore, the multifaceted nature of misinformation and disinformation requires AI/ML to analyse items of a visual nature. As mentioned above, there is an increase in the spreading of visual disinformation, in the form of manipulated images, or manipulated videos (i.e., deepfakes). For this reason, new tools such as RevEye and Sensity.ai focused on visual disinformation, employ AI/ML to detect the manipulation process in items of this nature. Hence, we can expect that the tendency to build AI-centred tools will increase in the future, even though this will present several challenges related to AI biases, AI training, accuracy, and false positives [25], challenges that can perhaps be mitigated with co-creation.

From the outcomes of this study, we reckon that it is possible that in the near future we will see new market entrants who have developed along the lines of simultaneously increased co-creation and increased AI/ML. If that development comes to fruition, then there will be a similar increased need for similar development in the area of what is here called policy [24], in order to balance the nature of co-creative work with a developed AI.

This is a conceptual study based on experts' opinions. The natural next step would be to bring users into the picture, and examine their perceptions exploring for example how design qualities are categorised into product definitions, c.f e.g. [44], to deeper explore user needs and predict the near future directions for the product developments of anti-disinformation tools

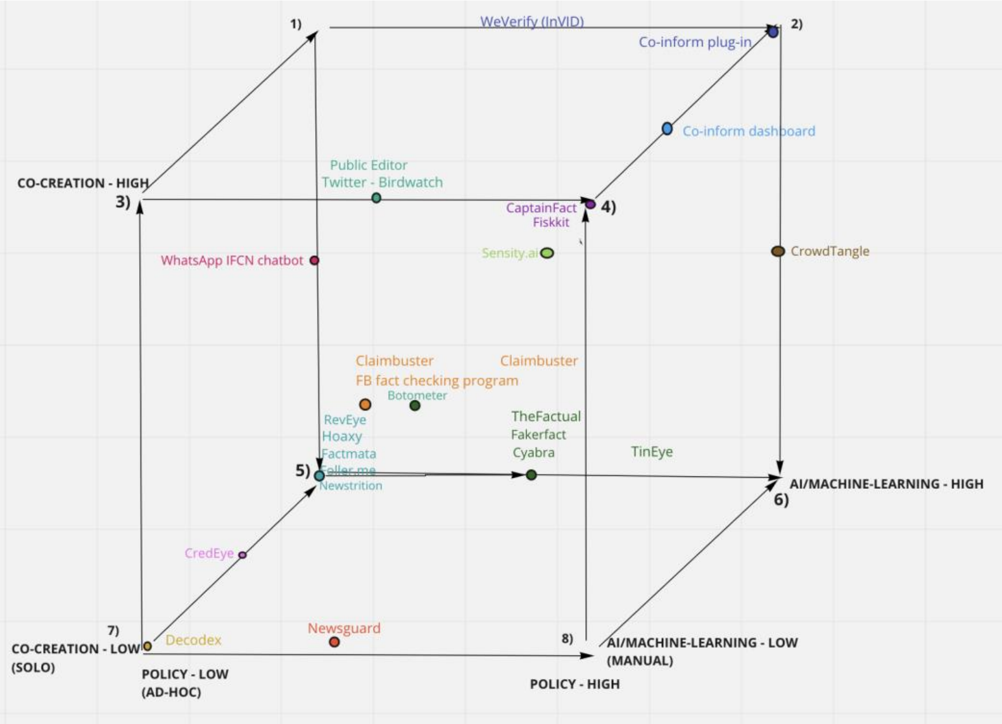

Figure 2 Distribution of the mapped tools across the cube

Appendix

Here follows a table that lists tools onto the three axes that make up the model. Some tools of these are in competition, some are complementary.

Table 1. Evaluation of the qualities of the selected items. For the three categories (i.e., policy, process, co-creation) each item is evaluated under three possible degrees, high, medium or low. The outcomes of this evaluation are reported visually in the cube.

Tool		Policy: (Re active vs Proactive) low/medium/high)	Process: (Manual vs AI/Machine Learning) (low/medium/high)	Creation: (Solo vs co creation) (low/medium/high)	Notes
Public Editor		medium	low	Community/high	
Botometer (by OSoMe)		low	high	low	
CaptainFact		high	low	high	Focus on inserting fact checks in videos.
Claimbuster		medium	medium	medium	Human assisted machine learning to identify potential claims
Co-Inform (plugin)		high	high	high	

Conform (dash board)		high	medium	high	
Cyabra		medium	high	low	
Decodex		low	low	low	Focus on source credibility

Facebook Fact- Checking Pro gram	medium	medium	medium	Third- Party Fact Checking Program with IFCN- verified partners
Factmata	low	high	low	
FakerFact (plugin)	medium	high	low	
Fiskkit	high	low	high	Works on the critical thinking and analysis skills
Foller.me	low	high	low	Similar to MisinfoMe
Hoaxy (by OSoMe)	low	high	low	Similar to MisinfoMe
NewsGuard (plugin)	medium	low	low	Focus on source credibility; run by journalists
Newstrition	low	high	low	
TheFactual	medium	high	low?	Focus on article credibility

TinEy.	medium	high	low	Reverse image search
Twitter Bird watch	medium	low	high	Crowd-sourced fact-checking
WeVerify (In VID) (plugin)	high	High	high	Focused on videos.
Whatsapp IFCN Chatbot	low	High	medium	
CredEye	low	medium	low	Web page. Produces credibility score of a given text
CrowdTangle	high	high	medium	
Sensity.ai	medium	high	medium	
RevEye	low	high	low	

References

1. Adikari, A., Burnett, D., Sedera, D., de Silva, D., & Alahakoon, D. (2021). Value co-creation for open innovation: An evidence-based study of the data driven paradigm of social media using machine learning. International Journal of Information Management Data Insights, 1(2), 100022.

2. Allport, G. W., & Postman, L. (1946). An analysis of rumor. Public opinion quarterly, 10(4), 501-517. https://doi.org/10.1093/poq/10.4.501

3. Ansari, A., Economides, N. and Steckel, J., 1998. The max-min-min principle of product differentiation. Journal of Regional Science, 38(2), pp. 207-230.

4. Babakar, M. (2018). Crowdsourced Fact checking. Retrieved February, 2021 from: https://medium.com/@meandvan/crowdsourced-factchecking-4c5168ea5ac

5. Bergmann, E. (2020). Populism and the politics of misinformation. Safundi, 21(3), 251-265. https://doi.org/10.1080/17533171.2020.1783086

6. Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. Data & Policy, 3.

7. Burgoon, J. K., Blair, J. P., Qin, T., & Nunamaker, J. F. (2003, June). Detecting deception through linguistic analysis. In International Conference on Intelligence and Security Informatics (pp. 91-101). Springer, Berlin, Heidelberg.

8. Burkhardt, J. M. (2017). History of fake news. Library Technology Reports, 53(8), 5-9.

9. Choy, M., & Chong, M. (2018). Seeing through misinformation: A framework for identifying fake online news. arXiv preprint arXiv:1804.03508.

10. Columbia Engineering. (2022-last update), Artificial Intelligence (AI) vs. Machine Learning. Available at: https://ai.engineering.columbia.edu/ai-vs-machine-learning/ [-10-18, 2022].

11. Deshmukh, A., & Wankhade, S. B. (2021). Deepfake Detection Approaches Using Deep Learning: A Systematic Review. Intelligent Computing and Networking, 293-302.

12. Di Domenico, G., Sit, J., Ishizaka, A., & Nunan, D. (2021). Fake news, social media and marketing: A sys tematic review. Journal of Business Research, 124, 329-341.

13. Ekström, M., Lewis, S. C., & Westlund, O. (2020). Epistemologies of digital journalism and the study of misinformation. New Media & Society, 22(2), 205–212. https://doi.org/10.1177/1461444819856914

14. Farrell, J., McConnell, K., & Brulle, R. (2019). Evidence-based strategies to combat scientific misinformation. Nature Climate Change, 9(3), 191-195. https://doi.org/10.1038/s41558-018-0368-6

15. Farrel, T., Mensio, M., Burrel, G., Picollo, L., & Alani, H. (2018). D3. 2 Survey of misinformation detection methods. Co-Inform Project.

16. Fernandez, M., & Alani, H. (2018, April). Online misinformation: Challenges and future directions. In Companion Proceedings of The Web Conference 2018 (pp. 595-602). https://doi.org/10.1145/3184558.3188730

17. Ford, E. (2012) What's in Your Filter Bubble? Or, How Has the Internet Censored You Today? Retrieved January, 2021 from http://pdxscholar.library.pdx.edu/cgi/viewcontent.cgi?article=1078&context=ulib_fac

18. FullFact, (2022-last update), Automated Fact Checking. Available: https://fullfact.org/about/automated/ [Feb 17, 2022].

19. Giglietto, F., Iannelli, L., Rossi, L., & Valeriani, A. (2016). Fakes, news and the election: A new taxonomy for the study of misleading information within the hybrid media system. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2878774

20. Gummesson, E., Mele, C., Polese, F., Galvagno, M. and Dalli, D., (2014). Theory of value co-creation: a systematic literature review. Managing Service Quality, .

21. Gunton, K. (2022). The Use of Artificial Intelligence in Content Moderation in Countering Violent Extremism on Social Media Platforms. In Artificial Intelligence and National Security (pp. 69-79). Springer, Cham.

22. Han, O., Baris, I., Hosseini, A. S., de Nigris, S., & Staab, S. (2019). Democratic policy-making for misinformation detection platforms by git-based principles. Human Computer Interaction and Emerging Technologies: Adjunct Proceedings from, 67.

23. Jaursch, J., Lenoir, T., Schafe, B. and Soula, E. (2019, November), Tackling Disinformation : Going Beyond Content Moderation. Available: https://www.institutmon taigne.org/en/blog/tackling-disinformation-going-beyond-content-moderation [Nov 10, 2020].

24. Jia, F. (2020). Misinformation Literature Review: Definitions, Taxonomy, and Models. International Journal of Social Science and Education Research, 3(12), 85-90.

25. Juršėnas, A., Karlauskas, K., Ledinauskas, E., Maskeliūnas, G., Rondomanskas, D., Ruseckas, J. (2022). The role of AI in the battle against disinformation. NATO Strategic Communications Centre of Excellence

26. Kertysova, K. (2018). Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. Security and Human Rights, 29(1-4), 55-81.

27. Komendantova, N., Ekenberg, L., Svahn, M., Larsson, A., Shah, S. I. H., Glinos, M., ... & Danielson, M. (2021). A value-driven approach to addressing misinformation in social media. Humanities and Social Sciences Communications, 8(1), 1-12. https://doi.org/10.1057/s41599-020-00702-9

28. Korshunov, P., & Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685.

29. Koulolias, V., Jonathan, G. M., Fernandez, M., & Sotirchos, D. (2018). Combating Misinformation: An ecosystem in co-creation. OECD Publishing.

30. Mahadevan A. (2021). Twitter's crowdsourced fact-checking experiment reveals problems. Retrieved from: https://www.poynter.org/fact-checking/2021/analysis-twitters crowdsourced-fact-checking experiment-reveals-problems/ [Mar 17, 2021].

31. Mensio, M. and Alani, H. (2019). News Source Credibility in the Eyes of Different Assessors. In: Conference for Truth and Trust Online, 4-5 Oct 2019, London, UK, (In Press).

32. Moon, K., & Blackman, D. (2017). A guide to ontology, epistemology, and philosophical perspectives for interdisciplinary researchers. Integration and Implementation Insights.

33. Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., ... & Martino, G. D. S. (2021). Automated fact-checking for assisting human fact-checkers. arXiv preprint arXiv:2103.07769.

34. Newman, N., (February, 2022), Journalism, media, and technology trends and predictions 2022. Retrieved from: https://reutersinstitute.politics.ox.ac.uk/journalism-media-and-technology-trends-and-predictions-2022 [Feb 17, 2022].

35. Osborne, P. S. (2018). From Public Service-Dominant Logic to Public Service Logic: Are Public Service Organizations Capable of Co-Production and Value Co-Creation?. Public Management Review 20 (2): 225–231. https://doi.org/10.1080/14719037.2017.1350461

36. Parisier, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin UK.

37. Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. Journal of experimental psychology: general, 147(12), 1865.

38. Pengnate, S. F. (2019). Shocking secret you won't believe! Emotional arousal in clickbait headlines. Online Information Review.

39. Posetti, J., & Matthews, A. (2018). A short guide to the history of 'fake news' and disinformation. International Center for Journalists, 7, 1-19.

40 Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: a systematic literature review. IEEE Access.

41. Sajithra, K., & Patil, R. (2013). Social media–history and components. Journal of Business and Management, 7(1), 69-74.

42. Schuster, T., Fisch, A., & Barzilay, R. (2021). Get your vitamin c! robust fact verification with contrastive evidence. arXiv preprint arXiv:2103.08541.

43. Seo, B. G., & Park, D. H. (2020). The effective type of information categorization in online curation service depending on psychological ownership. Sustainability, 12(8), 3321.

44.Svahn, M., & Lange, F. (2009). Marketing the category of pervasive games. In Pervasive games (pp. 219-230). Morgan Kaufmann.

45. Svahn, M. and Coppolini Perfumi, S., Sep 2021. A conceptual model for approaching the design of of anti-disinformation tools , Egove Cedem 2021, September 2021 Sep 2021, Springer.

46. Samoili, S., Cobo, M. L., Gomez, E., De Prato, G., Martinez-Plumed, F., & Delipetrev, B. (2020). AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence.

47.Tenove, C. (2020). Protecting Democracy from Disinformation: Normative Threats and Policy Responses. The International Journal of Press/Politics, 25(3), 517-537. https://doi.org/10.1177/1940161220918740

48. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131-148.

49. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151. https://doi.org/10.1126/science.aap9559 23.

50. Wardle, C. (2016). 6 types of misinformation circulated this election season. Columbia Journalism Review, 18.

51. Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking.

About the Authors

Mattias Svahn

Mattias Svahn is head of research at eGovlab, the Stockholm University Centre for Excellence in e-Governance Studies. Has has been a visiting professor at the Stockholm School of Economics in Riga and head of research for the Horizon 2020 project Co-inform. His research interests include, online disinformation, media psychology, and guerilla marketing.

Serena Coppolino Perfumi

Serena Coppolini Perfumi is a PhD student at the department of sociology at Stockholm University, Sweden. She has been WP-lead for WP:s in the Horizon 2020 project Co-inform. Her research interests include online disinformation, Internet cultures, and activism on social media platforms.