Addressing contingency in algorithmic misinformation detection: Toward a responsible innovation agenda

Andrés Domínguez Hernández*[1] ORCID 0000-0001-7492-7923, Richard Owen[2] ORCID 0000-0002-1767-3901, Dan Saattrup Nielsen[3] ORCID 0000-0001-9227-1470, Ryan McConville[3] ORCID 0000-0002-7708-3110

**Corresponding author: andres.dominguez@bristol.ac.uk*

[1] Department of Computer Science, University of Bristol

[2] School of Management, University of Bristol

[3] Department of Engineering Mathematics, University of Bristol

Abstract

Machine learning (ML) enabled classification models are becoming increasingly popular for tackling the sheer volume and speed of online misinformation. In building these models, data scientists need to take a stance on the legitimacy, authoritativeness and objectivity of the sources of 'truth' used for model training and testing. This has political, ethical and epistemic implications which are rarely addressed in technical papers. Despite (and due to) their reported high performance, ML-driven moderation systems have the potential to shape online public debate and create downstream negative impacts such as undue censorship and reinforcing false beliefs. This article reports on a responsible innovation (RI) inflected collaboration at the intersection of social studies of science and data science. We identify a series of algorithmic contingencies –key moments during model development which could lead to different future outcomes, uncertainty and harmful effects. We conclude by offering an agenda of reflexivity and responsible development of ML tools for combating misinformation.

Keywords and Phrases: misinformation, reflexivity, content moderation, fact-checking, machine learning, responsible innovation

Introduction

In recent years there has been a flurry of research on the automated detection of misinformation using Machine Learning (ML) techniques. Significant progress has been made on developing ML models for the identification, early detection and management of online misinformation[1], which can then be deployed at scale to assist human moderators (e.g., Hassan et al. 2017; Monti et al. 2019; Zhou, Wu, and Zafarani 2020). The development of these tools has gained currency particularly among social media platforms like Facebook, Twitter and YouTube given their key role in the propagation of online misinformation and mounting regulatory pressure to manage the problem. In response to the overwhelming scale of misinformation –notably in the context of the COVID-19 pandemic- and the limited capacity of human moderation to address this, platforms have increasingly looked to the deployment of automated models as standalone solutions requiring less or no human intervention (CDEI 2021).

[1] Several terms related to misinformation (e.g., disinformation and fake news) are used throughout this paper to refer to specific attempts in the literature to define and tackle related problems. However, the term misinformation, in its broadest sense, is preferred for analysis as it encompasses any type of misleading or false content presented as factual, regardless of intent.

The artificial intelligence (AI) research community has broadly framed the problem as one that can be tackled using ML–enabled classification models. These classify, with varying levels of accuracy, the category to which a piece of data belongs (e.g. 'factually true', 'false' or 'misleading' claim). These models are trained on large datasets of various modalities (images, text or social connections) containing properly annotated samples of information labelled as being factually correct or false by the model developer (Torabi Asr and Taboada 2019). In order to advance the state of the art, researchers strongly emphasize the need for more and higher quality data which can be used to train and validate ML models. Several training datasets have been published to this end containing collections of fake news articles, social media posts, fabricated images, or false claims along with labels about their truthfulness produced by fact-checking organizations around the world[2].

Recent work in fair-ML and critical data studies has started to examine the assumptions and practices surrounding the curation of training sets and their use in the construction of ML models. Of note are discussions relating to ethical issues of algorithmic discrimination, bias and unfairness (e.g., Binns et al. 2017; Cooper 2020; Jaton 2021; Miceli et al. 2021). However, scant attention has been given to the epistemic assumptions that underlie ML- enabled models for misinformation identification and, associated with this, their social, political and commercial entanglements. A persistent rationale in developing such tools is that if fed with (large and good) enough data they will

[2] A non-exhaustive list of datasets is found on https://data.4tu.nl/articles/dataset/Repository_of_fake_news_detection_datasets/14151755

be able to produce reliable, actionable evaluations of truthfulness, allowing users to tackle the problem of misinformation in an automated, cost-effective manner. The construction of referential datasets (or ‘ground truths’) used for both model training and performance measuring purposes is rooted in assumptions about the credibility and trustworthiness of these data sources. These sources typically include corpora of authoritative knowledge or the outputs of professional fact-checking organizations, which are implicitly assumed to be credible and can be used to benchmark what is ‘true’ –or at least not false. Explanations about what counts as ‘authoritative’ or ‘reliable’ ground truths and reflection on associated assumptions, limitations and ethical implications are rarely seen in technical papers describing model development and application.

This paper attempts to address this gap. We present findings from a responsible innovation (RI) inflected collaboration between science and technology studies (STS) scholars and data scientists developing ML-enabled tools to combat misinformation. This study was undertaken as part of a cross-cutting RI workstream exploring the integration of RI frameworks [references anonymised for peer review] within a large interdisciplinary research Centre [anonymised for peer review]. Our collaboration focused on mobilising the reflexivity (first and second order) and anticipatory dimensions of RI [author date]. Drawing on insights from feminist epistemology and social studies of science and expertise, we put forward a series of *algorithmic contingencies* –key moments during model development which could lead to different future outcomes, uncertainties and even harmful effects—and propose a responsible innovation agenda for ML- enabled misinformation detection and management. The frame of contingencies departs from the calculus of fairness or data bias elimination discussed in the literature to date (Selbst et al. 2019). We advance that taking these contingencies seriously opens a space for reflection, debate and the evaluation of the social value and potential harmful impacts of these tools.

Theoretical lens: On the social construction of facts, facticity and fact-checking

Who gets to decide whether a conjecture or claim meets the quality and condition of being a fact –i.e., establishes its facticity - is a contentious and contested matter. Reducing the establishment of facticity to 'checking' and 'verifying' loses the richness and complexity of fact construction as an inherently social process (Latour and Woolgar, 1986). As scholars within feminist epistemology, philosophy of science and studies of science and expertise have compellingly argued, facts do not exist in a value-free vacuum: they are crafted, contested and pondered against competing claims as they move within social worlds. Facts are thus necessarily contingent to context, cultural norms, institutional structures and power relations (Collins and Evans 2002; Haraway 2013; Jasanoff 2004; Latour and Woolgar 1986). Not only this, but over time the social and historical circumstances on which the construction of a fact depends can become opaque and lost; seemingly 'free from the circumstances of its production' (Latour and Woolgar, 1986 p103). In this shifting and contingent knowledge arena, the legitimacy of those who warrant and assert claims and conjectures becomes key.

Legitimacy can be granted through credentialled expertise, reputation and social acceptance (Yearley 1999). Relying on the authority of scientific expertise and reputable journalism could well be a socially acceptable means for determining what one might call ‘objective knowledge’. Feminist scholars have however challenged the objectivist ideal of ‘science as neutral’ as it paradoxically elides the forces that often shape knowledge production –Western, male, and elite dominated funding institutions, research priorities, special interest groups, etc. (Haraway 2013; Harding 1995). This observation is not a relativist attack on expertise and science as an institution, but a call to remain cautious about the often loose use of the language of ‘neutrality’ and ‘objectivity’ (Harding 1995; Lynch 2017). We take this caution as our starting point to examine how assumptions about (scientific) knowledge, expertise and facticity might become encoded in AI techniques aimed at sorting out and managing (mis)information.

While there are different computational approaches to combat misinformation, in this article we focus primarily on efforts to leverage and automate the journalistic practice of fact-checking through ML-based techniques (Hassan et al. 2017). In the last decade, professional fact-checking has gained prominence for its role in promoting truth in public discourse, especially during times of elections and crises (e.g., wars and pandemics). To date, there exist hundreds of professional fact checkers around the world[3]. Modern data-driven fact-checking has increasingly been viewed as vital to tackle misinformation in the

[3] A database of global fact-checking websites has identified more than 300, https://reporterslab.org/fact-checking/

so called ‘post truth’ era (Carlson 2017). Social media platforms, and notably Facebook, have partnered with professional fact-checkers around the world to combat the widespread misinformation problem[4]. Not only are fact-checkers entrusted with the moderation of dubious pieces of information flagged as such by platforms’ algorithms, but their verdicts are used to help train misinformation detection algorithms and ML models (CDEI 2021). The output and credibility of professional fact-checking is usually taken at face value for these purposes. However, as a human activity, professional fact-checking is not immune to cognitive and selection biases, subjectivity and ideological preferences, errors, and (geo)political and commercial interests. Despite being presented as impartial and objective, fact-checkers we suggest are political actors engaging in epistemic practices, i.e., establishing facticity and confronting lies (defined by them) in public discourse.

Fact-checking services have attracted some criticism over their methodologies due to, for example, accusations of skewed selection of topics, actors, and claims; and the use of ambiguous terminology[5] (Uscinski and Butler 2013; Stewart 2021). These issues manifest in myriad ways; for instance as competing or contentious verdicts between fact-checkers or shifting assessments of claims over time, sometimes with serious consequences (Lim 2018; Nieminen and Sankari 2021). Deceitful content and tactics are always evolving, but also, what constitutes a seemingly stable fact at a given point may

[4] See Facebook AI: ‘Here’s how we’re using AI to help detect misinformation’. November 19, 2020, https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/

[5] Fact-checking organizations often use vague, or borderline terminology (e.g., ‘mostly true’, ‘mostly false’) on the basis that claims are not always verifiable nor simply true or false.

change over time, driven by public debate or the availability of new information (Marres 2018). The COVID-19 lab leak controversy is a case in point. For the most part of 2020 the claim that COVID-19 originated in a lab in Wuhan, China, was widely dismissed by Western media as a conspiracy and 'fake news'. Early in 2021, growing calls to take the hypothesis seriously triggered further investigations by the WHO and a swift change of narrative by fact-checkers and the media (Thacker 2021). Amidst the controversy, Facebook automatically mislabelled a news article critical of the WHO as 'misinformation', which was later corrected after complaints of censorship by the news outlet[6]. This episode shows that while well intentioned, the practice of fact-checking can also lead to ambivalences and false positives which could in turn be blindly reproduced by an algorithm.

Not only the online misinformation ecology evolves quickly, but the experiences and manifestations of misinformation differ vastly across cultures, idiosyncrasies, languages and political realities (Prasad 2021; Seifert 2018). These ambiguities are not trivial for the design of interventions, as research has shown that the publication of fact-checks can have uneven effects on different audiences, depending on a person's beliefs or initial stance on the topic (Park et al. 2021; Walter et al. 2020). Furthermore, correction efforts could have the backfiring effect of reinforcing entrenched beliefs and the spread of

[6] See https://unherd.com/thepost/facebook-censors-award-winning-journalist-for-criticising-the-who/

misinformation due to the segregating dynamics of knowledge communities online formed around shared politics and identities (Nyhan and Reifler 2010; see also section 5). In pointing out the challenges involved with 'establishing the truth' we do not seek to undermine the value of expertise and journalism in public discourse. Albeit inevitably partial and context-dependent, truth-seeking efforts such as fact-checking can still be of use in the fight against misinformation. However, we contend that these practices and their normative claims warrant reflection and scrutiny, particularly as they get scaled up and automated.

Empirics and methodology

This article is the result of an interdisciplinary collaboration between data scientists leading a research project [omitted for peer review] focused on the use of machine learning for tackling online misinformation, and social scientists affiliated with the field of science and technology studies (STS). The collaboration was led by the social scientists (authors 1 and 2) as part of their cross-cutting work on responsible innovation within a major academic centre [omitted for peer review] hosted by their university. Our analysis is informed by regular meetings within the research team over a period of approximately 8 months and empirically grounded in the development of a ML system to detect online misinformation over that same period. The technical project was conducted by a team of data scientists (authors 3 and 4) and comprised a multimodal machine learning based study of misinformation on social media and, the development of an ML-enabled tool for misinformation detection and management. One of the outcomes of the data science project was a ‘misinformation dataset’ [authors date] which intends to capture the diverse ways in which misinformation manifests on social media. This dataset contains roughly 13,000 claims (of which 95% are labelled as misinformation) from 115 fact-checking organizations and, more than 20 million posts (‘tweets’) from the Twitter platform related to these claims. Aside from capturing a sizeable amount of the social media interactions associated with the claims, the dataset covers 41 languages and spans dozens of different events (e.g., COVID-19, Israel-Palestine conflicts) appearing on the platform over the course of a decade.

This work builds on the longstanding tradition in STS of opening the world of scientists and black-boxed technical systems to scrutiny through ethnographic accounts (Latour and Woolgar 1986; Pollock and Williams 2010) and extends previous efforts to integrate social and ethical considerations into processes of research and development (Schuurbiers 2011; Fisher, Mahajan, and Mitcham 2006). While ethnography has been the archetypical tool of STS theory and intervention, in this study we explicitly adopted a collaborative version of the method by shifting from an ethnographer/informer arrangement toward a joint endeavour between social scientists and data scientists (c.f. Forsythe 1993; Downey and Zuiderent-Jerak 2021). Collaborative ethnography can be viewed as ‘an approach to ethnography that *deliberately* and *explicitly* emphasizes collaboration at every point in the ethnographic process, without veiling it—from project conceptualization, to fieldwork, and, especially, through the writing process. Collaborative ethnography invites commentary from our consultants and seeks to make that commentary overtly part of the ethnographic text as it develops' (Lassiter 2005, 16). Our aim with this approach is not only to enrich the process of interpretation from field observations and qualitative analysis of technical work into writing, but to advance collective reflection and the co-development of ethical and responsible practices.

The discussions were aimed primarily at developing a schematization of the process of development of the ML detection model and the curation of the training datasets used to support this (see next section). We focused on the technical development phase of the project while concurrently analysing comparable works in the literature on automatic misinformation detection. Given the collaborative nature of this work, the analysis combines interpretative description and self-critical reflections co-produced by the research team. We approached this method iteratively by purposely surfacing the technical and epistemic assumptions and practices in ML model development for analysis. Subsequently, interpretative texts written by the social scientists were checked and expanded by the data scientists. We acknowledge the limitations of our method in producing generalizing claims which are reflective of our own concerns, experiences, the practices of a specific project and a limited subset of works in the literature. The next two sections describe the findings of our study.

Contingencies of automatic misinformation detection

In this section we describe the steps taken in the data science project to construct the automated misinformation model. We use this as a tangible way to critically examine the assumptions and practices involved in the development of ML tools for online misinformation detection. Below we first describe the generic model construction process adopted in the project which includes the steps of problem definition, choice of variables, curation of ground truth datasets, model validation and finally deployment (Figure 1). We then employ the notion of contingencies to interrogate the entanglements associated with the model, and the conditions that could alter the actual or claimed utility of a model and lead to deleterious consequences. Harmful impacts could for example include legitimate information being wrongly categorized as misinformation (false positives) and subsequently leading to unfair censorship; the amplification of objectionable or ambiguous truth assessments; or the reinforcing of false beliefs by failing to identify misinformation (false negatives). It is important to note that here we do not view biases as inherently negative; in fact, they could be necessary for the purposes of tackling misinformation. For instance, using expert sources such as scientists or reputable institutions to correct misinformation is a form of socially acceptable bias which may prove effective even though experts are fallible and may not always reach consensus on what is true of even what constitutes as being a fact (Latour and Woolgar 1986). Thus, this exercise is not aimed at debiasing or showing how to better locate facts and determine ‘truth’; in fact, it highlights the difficulties of doing so via manual or computational approaches. We illustrate the salience of various contingencies so that they are pondered reflexively in the development and audit of tools. In the following subsections we examine these contingencies in more detail, describing them first and then locating each more specifically in the context of the project's research and development.

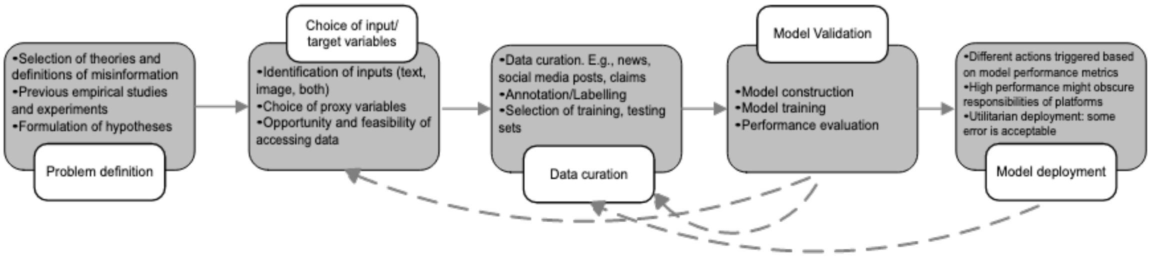
•Selection of theories and definitions of misinformation
•Previous empirical studies and experiments
•Formulation of hypotheses
Problem definition
Choice of input/ target variables
•Identification of inputs (text, image, both)
•Choice of proxy variables
•Opportunity and feasibility of accessing data
•Data curation. E.g., news, social media posts, claims
•Annotation/Labelling
•Selection of training, testing sets
Data curation
Model Validation
•Model construction
•Model training
•Performance evaluation
•Different actions triggered based on model performance metrics
•High performance might obscure responsibilities of platforms
•Utilitarian deployment: some error is acceptable
Model deployment

Figure 1: Steps in the construction of an automatic misinformation detection model. (1) Problem definition: design of strategy based on hypotheses, definitions and theories about how to identify misinformation. (2) Selection of multimodal inputs and target variables to be included into the classification model based on feasibility and opportunity. (3) Curation and structuring of the data and selection of training and testing subsets. (4) Model validation: ground truth datasets are used for training the model and testing its performance. Metrics of performance accompany the publication of classification models. (5) Model deployment: model outputs will trigger interventions such as banning, downranking or flagging misinformation. Performance metrics imply some misclassifications are tolerable.

Problem definition

The way misinformation is problematized shapes the strategy used to detect and identify it as well as the required data required and their structure. Within the AI/ML research community, misinformation detection is generally framed as a task of *classification* whereby candidate pieces of information are classified by a ML model according to sensible (i.e. legitimate) evaluations of their truthfulness[7]. These evaluations are influenced and informed by existing empirical studies, theories and hypotheses about what may constitute or signal the presence of misinformation. Misinformation and truth in this sense are defined a-priori by the researchers and then translated (*formalized*) into a ML model following different data-driven strategies which could leverage e.g., a corpus of knowledge, writing content and style or patterns of propagation (Gradoń et al. 2021). For example, 'fake news'–a popularized idiom in the misinformation landscape–is typically defined as a type of misinformation which is intentionally crafted, often with a political or financial interest. Based on that hypothesis different indicators such as style of writing or other distinctive features could be leveraged to single out and identify fake news from the analysed content (e.g., social media). For example, Rashkin et al. (2017) developed a model that classifies political statements and news based on *linguistic features* such as keywords or subjective language that indicates signs of intent to deceive. In that case, the authors drew on previous empirical work, communication theory and hypotheses suggesting that 'fake news articles try to enliven stories to attract readers'. Techniques of natural language processing are increasingly being used to support this. Another common strategy is to search for signs of misinformation (irrespective of intent) by looking at its *impact,* particularly regarding what is distinctive about information consumption patterns in relation to those exhibited by legitimate, 'truthful' information. For instance, several

[7] Classification models in this domain are typically *supervised*, that is, they require properly annotated data for training and testing.

studies have shown that false information online (e.g., fake news) tends to spread faster than verified information (Vosoughi, Roy, and Aral 2018). A model informed by such findings would rely on the assumption that social media metrics –such as likes, retweets or comments— can reveal something about the patterns of consumption of falsehoods (Monti et al. 2019).

In the [project name removed for peer review], multiple hypotheses underpinned the construction of the dataset which draw on previous studies, plausible assumptions or experiments conducted by the researchers. These were (1) people interact differently with posts discussing misinformation compared to posts discussing factually true information, in the sense of their replies and retweets (Shu et al. 2020; Li et al. 2020a); (2) the images used when discussing misinformation are different to images used when discussing factually true information (Jin et al. 2017); (3) users who are discussing misinformation tend to be different to those discussing factually true information as adjudged by their followers, followees and posts (Dou et al. 2021); (4) misinformation spreads faster on social media than factually true claims (Vosoughi, Roy, and Aral 2018); (5) posts discussing misinformation tend to use different hashtags to posts discussing factually true claims, where a hashtag is assumed to be 'different' from another hashtag depending on how it is used in tweets (Cui and Lee 2020; Li et al. 2020b); and (6) a classifier trained on data which is monolingual or monotopical will not generalize to new languages and events

(Han, Karunasekera, and Leckie 2020). The variables related to these hypotheses were used as a basis to construct a dataset so as to make those variables available for analysis.

Choice of inputs/target variables

The choice of inputs to a ML model not only reflects the designer's framing of the problem but the technical feasibility in terms of what kind of data can be reliably and economically acquired and used at scale. Most algorithmic techniques to date have used text as an input variable – however misinformation is often also contained in images, video, or can result from deliberately mismatched combinations of text and images intended to deceive or lure users; hence the growing focus on working with multimodal data, which is also salient in the [project name removed for peer review].

The lack of data is frequently mentioned in the literature as one of the biggest obstacles in the detection of misinformation. This is a twofold issue. On the one hand, some social networks make the acquiring of data more difficult (e.g., Facebook) while others make this easier (e.g., Twitter where academic access is 10 million tweets per month, and it is straightforward to get access). On the other hand, in order to use supervised ML models, online content needs to be annotated e.g., as 'true' or 'false' (see subsection 4.3). To get around the problem of general data scarcity, different forms of triangulation or proxy data are often included in the mix of building blocks. It may, for example, be relatively inexpensive to scrape social media for reactions or social engagement metrics which could be used to detect the presence of misinformation. For example, Lee et al., (2020) use sentiment analysis to quantify the ‘perplexity’ users express in their comments to social media posts as a proxy of early signs of misinformation. Because some models rely on statistical correlations between variables in lieu of causal relations, there is the risk of them making spurious associations. For instance, using data from satirical news as a source to identify fake news could lead to wrong associations due to the presence of confounding variables such as humor (Pérez-Rosas et al. 2018).

Opportunity, resources and feasibility were key considerations within the [project name removed for peer review]. The strategy was to use a *feature-rich* dataset containing as many modalities as possible with the expectation that the combination of these could lead to better (i.e. more accurate) predictions of misinformation. Given the problem was formulated as a binary classification task, the target variables were defined as ‘factual’ and ‘misinformation’. The model relied on annotated, multimodal data (i.e., texts and images) obtained from the Google Fact Check API[8] as well as from Twitter which provides access to data about user engagements with news. This was motivated by the ready availability of data for research purposes and because the widespread use of these sources in the automatic misinformation detection research literature would make it easier to test whether the current effort would *outperform* existing models.

[8] This service aggregates claims which have been fact-checked by eligible news organisations. To be included in Google’s fact check tool, news organizations need to comply with Google’s standards for publishers and content policies. https://support.google.com/news/publisher-center/answer/6204050

Data curation

In order to train a classification model, a labelled dataset –or ground truth—is needed as a referent of factual and non-factual information. Curating a ground truth set is not a trivial act but requires the machine learning developer to make principled choices as to what is an acceptable source of legitimate information. As discussed above, a common and defensible approach to determining ground truths is to defer to experts or authoritative sources of knowledge. For example, Wikipedia, reputable news outlets, professional fact-checkers and 'wisdom-of-the-crowd' have been used to build labelled datasets of categorized (mis)information. While using science-informed sources is seldom objectionable, there are still conditioning factors. For instance, in some circumstances, deference to experts may be unwarranted[9]; and journalists might (unintendedly or not) publish data in a way that is skewed and misleading (Lewis and Westlund 2015). For exposition, here we enumerate some of the conditionalities associated with ground truth datasets based on fact-checking.

First, ground truths are highly *contingent on timing* and thus *have diminishing returns.* This is because the online (mis)information environment is in constant flux. A model trained on previously fact-checked information is likely to be more effective with similar or comparable content and less so with whole new topics, themes and genres of false content. As the COVID-19 lab leak controversy demonstrates, factual assessments

[9] An illustration of this is what Rietdijk and Archer (2021) problematize as 'false balance' in journalism. This issue has been particularly salient in the debate around climate change, where journalists have given disproportionate attention to a minority of climate sceptics within the scientific community, who may also qualify as experts, in their efforts to show both sides of the debate.

could shift dramatically over a short period of time. These shifts are not always adequately and consistently addressed by fact-checkers[10] such that they can be taken onboard in updating an ML model. If left unchecked, ambivalences in the training data could lead to the amplification of objectionable results and potentially harmful false positives.

Another key conditionality of constructing ground truth datasets is the choice of labels and labelling systems. This is particularly problematic when truth assessments are expressed in ways that are *ambiguous or subject to multiple interpretations*. One of the biggest challenges with using the work of fact-checkers as a source of ground truths is the lack of consistency among fact-checkers' definitions, terminology and methodology, particularly in cases where misinformation is not blatant, but subtle and nuanced. Different organizations use different types of labelling, including politically charged phrasing ('pants on fire'), borderline ('mostly true', 'mostly false') or detailed assessments of claims when it comes to nuanced content which cannot be easily classified as either true or false. Such ambiguities inevitably demand data scientists to interpret, standardize or develop new labels from existing data. For instance, to tackle the issue with inconsistent labels in the [project name removed for peer review], the ML model was trained to classify the individual verdicts into three categories: 'factual', 'misinformation' and 'other'. The last category was included to handle verdicts which were not conclusive, such as 'not sure'

[10] For example, PolitiFact, a well-known fact-checker organization, decided to archive their original assessment on the lab leak controversy by removing it from their database and revising their assessment as 'widely disputed' (see https://www.politifact.com/li-meng-yan-fact-check/)

–the claims whose verdicts belonged to the ‘other’ category were not included in the final dataset. Training such a model requires labelled verdicts to be standardized even if this introduces new ambiguities and loss of nuance. For example, ‘Half true’ was categorized as ‘Misinformation’. To mitigate these ambiguities, a decision was made in the project to only include claims whose verdicts from the fact-checking organizations were unanimous.

Training data could also be *skewed toward false claims*. While fact-checkers attempt to validate true information and attempt to promote factual content, much of their work is focused on debunking falsehoods[11]. This is reflected in the composition of the ground truth datasets, for example, when they contain disproportionately more samples of falsehoods, or only one label for ‘fully true’ and several ones for dubious content ranging from ‘mostly true’ to ‘blatantly false’. This issue can lead to misrepresentation of truthful content (labelled as such) in datasets which undermines the ability of a model to accurately identify true statements (true negatives) and reduce false positives. The bias toward false claims can be viewed as a technical problem of unbalanced data which developers can attack by attempting to diversify content and sources in the construction of the dataset (Gravanis et al. 2019). However, balancing a dataset is not always a straightforward task. This was true in [project name removed for peer review] where the resulting dataset was largely skewed towards claims belonging to the ‘misinformation’ category (~95% of the claims). A choice was made to not balance the two categories by including, e.g., news

[11] Some fact-checking organizations focus exclusively on false and misleading claims (e.g., factcheck.org)

articles from ‘trusted sources’, as this would both introduce more bias as well as potentially *polluting* data from a different data distribution. In attempting to balance the data, ML models could be able to distinguish between new and old data, rather than distinguishing between factual and misleading claims, making the task superficially easier yet futile.

*Datasets bear human selection and cognitive biases***.** A crucial and difficult question for the practice of fact-checking is which claims are eligible for assessment. Fact-checkers necessarily incur selection biases when deciding which claims to check and which ones to leave out. This is particularly controversial in the assessment of political discourse where judgements are often passed on statements which may contain a mix of opinion and verifiable facts. According to Uscinsky (2015), one of the perils of fact-checking is the choice to assess ideologically charged claims or future predictions for factual accuracy even when these can only be verified retrospectively or are not verifiable at all. Similarly, selection biases might lead to uneven representation of content among fact-checking organizations. For instance, a comparative study of two major fact-checking organizations in the US found that not only did they rarely look at the same selection of statements but even when they did there was little agreement on how they scored ambiguous claims such as ‘mostly true’ or ‘mostly false’ (Lim 2018). Selection biases are not only a source of uncertainty, but they normatively influence what types of information are worthy of checking and which narratives are prioritized over others.

Model validation

The merit and utility of a classification model is judged by its ability to predict human generated labels. Once a model is trained, its accuracy can be measured by comparing the resulting classifications against an *unseen* subset of the ground truth data. For example, in the case of models using datasets with labels provided by fact-checkers (e.g., true or false), 100% accuracy on the test set will theoretically equate to the model correctly predicting all the labels given by the fact-checker on data not seen by the model during the training process.

This is a process of internal validation which is typically agnostic to how the model functions in the world and the possibility of downstream harmful impacts. Performance metrics (be they *accuracy*, *precision*, *recall*, and F_1*-score*[12]) are commonly used as indicators of relative incremental progress within the field and are used for comparisons against benchmarks of human decision-making or other competing algorithmic techniques. However, these comparisons may be decontextualized; that is, based on metrics alone without regard to the specific (thematic, temporal or cultural) domains in which different models were trained and the qualitative differences between them. Such decontextualization can be misleading as a model trained on e.g., political misinformation, is likely to be inadequate to detect misinformation in the celebrity domain (Han,

[12] Here accuracy is the proportion of the model's predictions which are correct, recall is the proportion of the positive samples which the model correctly predicted, precision is the proportion of the model's positive predictions which are correct. The F1-score is the harmonic mean of the recall and precision, which implies that if one of these two metrics are low then the F1-score will be correspondingly low as well.

Karunasekera, and Leckie 2020). Since accuracy metrics are not always indicators of good model performance they could be deceiving, particularly in models using imbalanced or unevenly represented datasets which still exhibit relatively high accuracy (Valverde-Albacete and Peláez-Moreno 2014). In the development of the [project name removed for peer review] dataset, diversity of the data was deemed of high priority, as existing benchmarking datasets are biased towards specific languages, topics or events. As the system sought to detect misinformation within unseen events, the dataset was not merely split at random into a training and testing part. Instead, these splits were created according to distinct events, thus making more consistent evaluations, albeit substantially harder. Further, the dataset was heavily unbalanced (95% of the data belongs to the misinformation category) which means an accuracy metric would not be very telling and therefore F_1-scores for the two categories were reported instead.

Despite their salient shortcomings, performance metrics have *performative power*[13] in that they create expectations around, and effectively vouch for, the value of an algorithm. Whether, and how, to deploy an ML model can be informed by various metrics of performance. For instance, if a model exhibits a relatively high level of accuracy in classifying fake content, this can be used as a justification for deploying a system without human moderation. According to Pérez-Rosas et al., (2017) models with over 70%

[13] The concept of language performativity is used here in the same sense as within language anthropology, gender studies and sociology of expectations. A claim or statement is thought of as performative insofar as it constitutes and *act* which has an effect in the world (see Borup et al. 2006; Hall 1999).

accuracy are generally considered as being as good as humans to identify fake news (to use the authors' term), yet they still have considerable room for errors. Metrics of accuracy, precision, recall and F_1-score not only provide an opportunity for granular performance evaluation, but they can crucially inform what specific actions can be triggered by a model. For example, a model with high precision (low false positive) and low recall (high false negative) may be deemed more useful in fully automated scenarios as, while it may miss many cases of misinformation, there will be more confidence that those it detects will be correct. On the other hand, in scenarios with human moderation, a model with lower precision, but higher recall, may be more useful as it will retrieve more possible misinformation than the former model, albeit at the expense of false positives, which can be corrected by human moderators.

Anticipating emergent issues during model deployment

There are several ways in which social media platforms implement detection algorithms. They can either configure hybrid decision-support systems (e.g., ML-assisted fact-checking) or operate as standalone, automated moderation systems with no human intervention. In the case of Facebook, content flagged by an algorithm as potentially false is typically relayed to independent fact-checkers who will make decisions on the veracity of the claim (CDEI 2021). This is a strenuous process requiring a great deal of manual input to process the huge amount of content circulating on social media.

Depending on the platform's policy, detection algorithms can trigger specific corrective actions such as banning/flagging/downranking content or promoting relevant verified information alongside deceitful posts (Gillespie 2020; Gorwa, Binns, and Katzenbach 2020). One of the pitfalls of such corrective approaches (and moderation policies at large) is that they are typically applied at global scale (affecting billions of people) with little regard to different demographics and socio-political contexts and in line with the company's (shareholders) values and definitions of what counts as being acceptable. Moreover, there is widespread evidence that major social media platforms have facilitated the formation of knowledge communities where content is circulated and segregated based on shared politics and interests (Cinelli et al. 2021; Sacco et al. 2021). Because of this, interventions relying on blanket corrective mechanisms may not only have disparate effects when used across different groups and cultural contexts but pose the risk of reinforcing false beliefs particularly amid those groups where the circulation of falsehoods or conspiracy theories is more prevalent (Nyhan and Reifler 2010). Existing algorithmic techniques still have limited ability to account for nuances in language, intent, cultural references, or sarcasm (Duarte, Llanso, and Loup 2018). This makes algorithms highly fragile when it comes to 'borderline' or tricky cases but also vulnerable to being circumvented by the creators of false content emulating the style of truthful sources or translating posts into other languages. Similarly, the overreliance on seemingly high performing algorithms risks worsening issues of unjustified censorship when content is wrongly identified as being false and is subsequently banned or downranked. Performance metrics are often invoked to frame a complex social problem within a logic of optimization. If errors are low, platforms tend to dismiss them as negligible or outweighed by the benefits of improved efficiency thereby shifting the burden of errors to an acceptable minority of affected users who are faced with appeal processes[14].

A perhaps more fundamental issue with the use of algorithmic misinformation detection is that it emphasizes the role of bad actors in the production and spread of misinformation at the expense of downplaying the interests and responsibilities of major technology companies. The business model of social media platforms is based on maximizing the time users spend on their platforms in order to generate advertising revenue. This is achieved through opaque algorithms of personalization and recommendation based on people's behaviour, demographics and preferences (Zuboff 2019). The attention economy rewards the circulation of (and engagement with) content regardless of its quality; and in fact, misinformation has been found to consistently receive widespread attention and engagement in social media platforms (Edelson et al. 2021). The commercial incentive of platforms to maximize engagement is thus at odds with the goal of meaningfully tackling the spread of misinformation and any type of harmful content. Furthermore, there is a risk that automation is positioned as a solution to the spread of

[14] As recently admitted by YouTube's representative: 'One of the decisions we made [at the beginning of the pandemic] when it came to machines who couldn't be as precise as humans, we were going to err on the side of making sure that our users were protected, even though that might have resulted in a slightly higher number of videos coming down.' (Neal Mohan quoted in Barker and Murphy 2020)

harmful content that ensures business as usual. While we recognize that automated detection can have promising benefits to deal with the scale of misinformation, its development should not overshadow broader debate around regulation and oversight of platforms.

Discussion: Toward a responsible innovation agenda

We contend that insofar as a model's outputs are underpinned by a series of contingent assumptions, institutional commitments and socially constructed assessments of facticity, there can be no such thing as an impartial or neutral (mis)information classifier. The contingencies associated with developing ML classification models evidence that multiple reasonable strategies and outcomes are possible and that these are necessarily influenced by the subjectivities and interests implicated in their development. Further, we emphasize that misinformation detection algorithms are highly temporally sensitive: models using historic data may quickly become obsolete and hence need to be routinely assessed considering up-to-date information and changing moderation norms set by platforms and regulatory bodies. A constructive question arising from the contingencies outlined here is what measures can be taken in the interest of harnessing the social value of algorithmic classification and minimizing any harmful effects. There is no straightforward procedure to establish what the *right* outcomes might look like given that desired outcomes, harmful effects and social preferences might be in conflict. For instance, while some might be in favour of reducing the volume of misinformation online by maximizing a model's true positives with a tolerable error, others will be disproportionally harmed by unfair censorship and undermined freedom of expression resulting from misclassifications. Similarly, some would argue that people have the right to share misinformation particularly if it is harmless, whereas potentially dangerous content could provide a justification for restrictions on freedom of expression. Yet in practice, drawing boundaries between harmful/harmless content and the limits to free expression is seldom a trivial exercise.

These are ongoing tensions which should not be rendered as solvable problems. Instead, the question of how we might produce socially beneficial ('good' or 'fair') algorithmic tools calls for careful attention to broader socio-technical, legal, political and epistemic considerations. We suggest developers should endeavour to account for algorithmic contingencies and reflect on the limitations of their creations. This implies a commitment to openness and self-critical reflection, making the assumptions and the various human choices throughout the stages of data curation, model construction and validation available for scrutiny and contestation by external observers and taking their potential for harmful outcomes seriously. While this is an open research challenge, we offer some practical recommendations aimed at developing a responsible innovation agenda in this field.

Reflexivity beyond datasets

Principled and defensible criteria such as relevance, authoritativeness, data structure and timelines of the truth assessments all provide a strong foundation for the curation of ground truth datasets. There are important ongoing efforts to improve the transparency of datasets which are of relevance here (Gebru et al. 2021; Geiger et al. 2020; Gilbert and Mintz 2019). However, we propose that accounting for contingencies, particularly in politically sensitive scenarios, requires going beyond considerations of data accuracy, reliability and quality to *acknowledge the complex processes of social construction* which configure the development and use of ML models. A recent study by Birhane et al., (2021) showed that highly cited ML research has typically ascribed to values of performance, efficiency and novelty over considerations of social needs, harms and limitations; yet most often researchers make implicit allusions to the value neutrality of research. Insofar as developers outsource the assessments of facticity to other actors and select particular topics or events as matters of concern, it becomes more crucial to examine one's own assumptions, biases and commitments which directly influence model development and that these are made available for auditing purposes. Misinformation classification is by necessity a value-laden practice with profound normative implications concerning the validity, quality, representativeness and legitimacy of knowledge. Linking back to the efforts of feminist scholars in surfacing the politics of knowledge production, we are reminded of the need to reject 'view from nowhere' ideals and practice reflexivity (Harding 1995; Suchman 2002). In the interpretative research tradition, reflexivity has been a standard of academic rigor and credibility which is attained through acknowledging prior biases, positionalities, experiences and prejudices impacting researchers' claims to knowledge. There is no reason why improving the credibility of scientific endeavours through reflexivity should not extend to the development of machine learning models. Indeed, reflexivity, a key concept in the RI literature [author date], has begun to be invoked in numerous calls for more transparency and accountability in the field of data science at large (D'Ignazio and Klein 2020; Tanweer et al. 2021; Miceli et al. 2021). There remain a great deal of practical challenges with attaining the intended virtues of reflexivity in organisational spaces fraught with multiple, conflicting logics such as universities [author date]. Despite this, we support a reflexive turn in data science and recommend much needed further research in this direction. At the very least, a reflexive and transparent approach seeks to avoid shifting the blame to the data and external sources, acknowledge partiality (as opposed to deceptive efforts to debias) and the distribution of collective responsibility within the actors and institutions involved in constructing and deploying a model. In order to surface algorithmic contingencies, we suggest transparency reports need to be complemented with reflexive disclaimers about developers' methodological choices, problem statements, institutional affiliations and sources of funding influencing data collection and model construction.

Situated and timely evaluations

Mechanisms should be in place to adjust the behaviour of a model or even the decision as to whether to deploy a model or not with regards to changing circumstances, information or situated community norms. For instance, changes in the terms of service of a platform, or relevant local norms and regulations (e.g., GDPR) should be taken into account along with dataset labels changing as a result of new information (e.g., fact checkers changing their original decision). Equally, if a user deletes their post, this should be removed from the training / test set. Thus, datasets, even if they do not collect any more data, do not remain static – in fact they can decrease in size over time – as the training and test data changes, the model performance will change. This would allow for some form of a dynamic and adaptive environment, where published models and their results are continuously re-evaluated. Community benchmarks based on location, language and domain-relevant test sets is one way to encourage this. These evaluations should be consistent so that model comparisons are made with attention to topicality, timing, context, language or different modalities used. Benchmark tests, for instance, could be conducted against models trained on data labelled by different fact-checkers to investigate the impact of potential political, selection or cognitive biases in the outcomes of a model. Further research is needed around how conditioning factors such as fact-checkers' political leanings, domain specialisms and location could be factored into the quantitative or qualitative algorithmic evaluations.

Accounting for and communicating uncertainty

Epistemic uncertainty in machine learning extends to the lack of knowledge about the outputs of a model or ignorance on the part of the decision maker. This can reflect subconscious biases, inaccuracies or gaps in the data as well as forms of data reduction or standardization of labels carried out by developers. While the temporary fix of omitting ambiguous verdicts (such as ‘half true’ or ‘mostly false’) might reduce the burden for moderators and (superficially) increase accuracy, it comes at the cost of unwanted outcomes such as casting doubt on legitimate information, doing away with important nuances in language, or leaving subtle misleading claims unaddressed.

Model construction is also impacted by aleatoric uncertainty. The data collection process from a social media platform can be stochastic for various reasons –Twitter for example provides only a sample of the stream of online posts for searches (thus two people searching for ‘coronavirus’ using that API may collect different results and thus create a different dataset). Even if not using this, arbitrary decisions (such as only keeping a subset of the available data) made throughout the data collection process are often not completely documented and may thus be irreproducible. Thus, the data collection systems themselves should be made public and available. Even if not fully reproducible (due to the dynamic nature of the social media platform, or their stochastic APIs), developers should provide complete executable documentation on the data collection process. This approach is considered as a way forward in the [project name removed for peer review]: the data collection platform has been released on GitHub[15] so others can see and execute the exact code used to build the dataset, and thus all decisions that were made.

[15] [Place holder for a link to dataset]

Acknowledging and responding to uncertainty is a key aspect of responsible innovation [author date]. Measuring and reporting the uncertainty of a model can be crucial in aiding human intervention and increasing the transparency of the system. Measures of uncertainty could be published in tandem with other metrics as part of the model evaluation. This could help to avoid overreliance on algorithms and minimize ambiguous outcomes by helping human moderators but also to ponder alternative interventions such as adding context to ambiguous content or links to contrasting news.

Conclusion

Through an interdisciplinary RI collaboration we have collectively and critically reflected on emerging efforts to identify and manage online misinformation at scale using machine learning classification models. In doing so we have proposed an agenda for a more reflexive and responsible development of these tools. The development and widespread use of automated misinformation detection systems raise pressing political, epistemic and ethical issues. We argue that, albeit promissory developments, these tools are highly contingent to the epistemic status of their ground truths, and the assumptions, choices and definitions underpinning their development, the contexts within which they are deployed and the interests of powerful actors in vetting the circulation of information online. We laid out a series of contingencies across the different stages in the construction of these models and assessed how assumptions of expertise and legitimacy, ideological biases, and commercial and (geo)political interests may influence the normative outcomes of models which are predicated as robust, accurate and high performing. We note that while our analysis is grounded on a specific issue tackled by ML, similar concerns are likely to hold true in other areas, particularly in the moderation of hate speech and terrorist content. This marks paths of future inquiry where our analytical approach could be used to interrogate the epistemologies and forces driving other algorithmic systems.

This study exemplifies an attempt to integrate RI in a project within a major research centre, bringing social sciences methods and theory into conversation with those of data science. We hope our contribution sparks further exploration of how data scientists and social scientists can work together so as to break with traditional and perhaps unproductive divisions of labour when research questions seem to fall out of the remit of one discipline or the other (Moats and Seaver 2019; Sloane and Moss 2019). We approached this study in an experimental fashion and as part of a single research team conducting iterative cycles of observation, interpretation, validation and calibration. Given that the present study was conducted in tandem with the development of an ML tool, our proposed agenda calls for further experimentation and testing in practice. While some of the contingencies systematically outlined here were well understood beforehand and duly considered in the project's technical pipeline either through technical adjustments or disclaimers in documentation, others were not immediately apparent and could only be attended to retrospectively. This flagged to us the value of engaging in collaboration early in the stages of development, a recommendation with which we close.

Disclosure Statement

No potential conflict of interest was reported by the author(s)

Funding

The research for this project was supported by REPHRAIN: The National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online, under UKRI grant: EP/V011189/1

Ethics declaration

This study was designed as an internal research project and approved by the ethics committee of the authors' institution following fair data management practices, informed consent and responsible research and innovation considerations.

REFERENCES

Barker, Alex, and Hannah Murphy. 2020. 'YouTube Reverts to Human Moderators in Fight against Misinformation'. *Financial Times*, 20 September 2020. https://www.ft.com/content/e54737c5-8488-4e66-b087-d1ad426ac9fa.

Binns, Reuben, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. 'Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation'. In *Social Informatics*, edited by Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri, 405–15. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-67256-4_32.

Birhane, Abeba, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. ‘The Values Encoded in Machine Learning Research’. *ArXiv:2106.15590 [Cs]*, June. http://arxiv.org/abs/2106.15590.

Borup, Mads, Nik Brown, Kornelia Konrad, and Harro Van Lente. 2006. ‘The Sociology of Expectations in Science and Technology’. *Technology Analysis & Strategic Management* 18 (3–4): 285–98. https://doi.org/10.1080/09537320600777002.

Carlson, Matt. 2017. *Journalistic Authority: Legitimating News in the Digital Era*. Columbia University Press.

CDEI. 2021. ‘The Role of AI in Addressing Misinformationon Social Media Platforms’. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1008700/Misinformation_forum_write_up__August_2021__-_web_accessible.pdf.

Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. ‘The Echo Chamber Effect on Social Media’. *Proceedings of the National Academy of Sciences* 118 (9). https://doi.org/10.1073/pnas.2023301118.

Collins, H.M., and Robert Evans. 2002. ‘The Third Wave of Science Studies: Studies of Expertise and Experience’. *Social Studies of Science* 32 (2): 235–96. https://doi.org/10.1177/0306312702032002003.

Cooper, A. Feder. 2020. 'Where Is the Normative Proof? Assumptions and Contradictions in ML Fairness Research'. *ArXiv:2010.10407 [Cs]*, November. http://arxiv.org/abs/2010.10407.

Cui, Limeng, and Dongwon Lee. 2020. 'CoAID: COVID-19 Healthcare Misinformation Dataset'. *ArXiv:2006.00885 [Cs]*, November. http://arxiv.org/abs/2006.00885.

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. MIT Press.

Dou, Yingtong, Kai Shu, Congying Xia, Philip S. Yu, and Lichao Sun. 2021. 'User Preference-Aware Fake News Detection'. *ArXiv:2104.12259 [Cs]*, April. http://arxiv.org/abs/2104.12259.

Downey, Gary Lee, and Teun Zuiderent-Jerak, eds. 2021. *Making & Doing: Activating STS through Knowledge Expression and Travel*. The MIT Press. https://doi.org/10.7551/mitpress/11310.001.0001.

Duarte, Natasha, Emma Llanso, and Anna Loup. 2018. 'Mixed Messages? The Limits of Automated Social Media Content Analysis'. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 106–106. PMLR. https://proceedings.mlr.press/v81/duarte18a.html.

Edelson, Laura, Minh-Kha Nguyen, Ian Goldstein, Oana Goga, Damon McCoy, and Tobias Lauinger. 2021. 'Understanding Engagement with U.S. (Mis)Information News Sources on Facebook'. In *Proceedings of the 21st ACM Internet*

Measurement Conference, 444–63. IMC ’21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3487552.3487859.

Fisher, Erik, Roop L. Mahajan, and Carl Mitcham. 2006. ‘Midstream Modulation of Technology: Governance From Within’. *Bulletin of Science, Technology & Society* 26 (6): 485–96. https://doi.org/10.1177/0270467606295402.

Forsythe, Diana E. 1993. ‘Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence’. *Social Studies of Science* 23 (3): 445–77. https://doi.org/10.1177/0306312793023003002.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. ‘Datasheets for Datasets’. *ArXiv:1803.09010 [Cs]*, December. http://arxiv.org/abs/1803.09010.

Geiger, R. Stuart, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. ‘Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?’ In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 325–36. FAT* ’20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372862.

Gilbert, Thomas Krendl, and Yonatan Mintz. 2019. ‘Epistemic Therapy for Bias in Automated Decision-Making’. In *Proceedings of the 2019 AAAI/ACM Conference* *on AI, Ethics, and Society*, 61–67. AIES '19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3306618.3314294.

Gillespie, Tarleton. 2020. 'Content Moderation, AI, and the Question of Scale'. *Big Data & Society* 7 (2): 2053951720943234. https://doi.org/10.1177/2053951720943234.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance'. *Big Data & Society* 7 (1): 2053951719897945. https://doi.org/10.1177/2053951719897945.

Gradoń, Kacper T, Janusz A. Hołyst, Wesley R Moy, Julian Sienkiewicz, and Krzysztof Suchecki. 2021. 'Countering Misinformation: A Multidisciplinary Approach'. *Big Data & Society* 8 (1): 20539517211013850. https://doi.org/10.1177/20539517211013848.

Gravanis, Georgios, Athena Vakali, Konstantinos Diamantaras, and Panagiotis Karadais. 2019. 'Behind the Cues: A Benchmarking Study for Fake News Detection'. *Expert Systems with Applications* 128 (August): 201–13. https://doi.org/10.1016/j.eswa.2019.03.036.

Hall, Kita. 1999. 'Performativity'. *Journal of Linguistic Anthropology* 9 (1/2): 184–87.

Han, Yi, Shanika Karunasekera, and Christopher Leckie. 2020. 'Graph Neural Networks with Continual Learning for Fake News Detection from Social Media'. *ArXiv:2007.03316 [Cs]*, August. http://arxiv.org/abs/2007.03316.

Haraway, Donna. 2013. ‘Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective’. In *Simians, Cyborgs, and Women*. https://doi.org/10.4324/9780203873106-18.

Harding, Sandra. 1995. ‘“Strong Objectivity”: A Response to the New Objectivity Question’. *Synthese* 104 (3): 331–49.

Hassan, Naeemul, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. ‘Toward Automated Fact-Checking: Detecting Check-Worthy Factual Claims by ClaimBuster’. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1803–12. KDD ’17. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3097983.3098131.

Jasanoff, Sheila. 2004. *States of Knowledge : The Co-Production of Science and the Social Order*. Routledge. https://doi.org/10.4324/9780203413845.

Jaton, Florian. 2021. ‘Assessing Biases, Relaxing Moralism: On Ground-Truthing Practices in Machine Learning Design and Application’. *Big Data & Society* 8 (1): 20539517211013570. https://doi.org/10.1177/20539517211013569.

Jin, Zhiwei, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. ‘Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs’. In *Proceedings of the 25th ACM International Conference on Multimedia*, 795–816.

MM ’17. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3123266.3123454.

Lassiter, Luke E. 2005. *The Chicago Guide to Collaborative Ethnography*. University of Chicago Press.

Latour, Bruno, and Steve Woolgar. 1986. *Laboratory Life: The Social Construction of Scientific Facts*. Sage Library of Social Research ; v.80. Beverly Hills: Sage Publications.

Lee, Nayeon, Yejin Bang, Andrea Madotto, and Pascale Fung. 2020. ‘Misinformation Has High Perplexity’. *ArXiv:2006.04666 [Cs]*, June. http://arxiv.org/abs/2006.04666.

Lewis, Seth C., and Oscar Westlund. 2015. ‘Big Data and Journalism: Epistemology, Expertise, Economics, and Ethics’. *Digital Journalism* 3 (3): 447–66. https://doi.org/10.1080/21670811.2014.976418.

Li, Yichuan, Bohan Jiang, Kai Shu, and Huan Liu. 2020a. ‘Toward A Multilingual and Multimodal Data Repository for COVID-19 Disinformation’. *2020 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/BigData50022.2020.9378472.

———. 2020b. ‘MM-COVID: A Multilingual and Multimodal Data Repository for Combating COVID-19 Disinformation’. *ArXiv:2011.04088 [Cs]*, November. http://arxiv.org/abs/2011.04088.

Lim, Chloe. 2018. ‘Checking How Fact-Checkers Check’. *Research & Politics* 5 (3): 2053168018786848. https://doi.org/10.1177/2053168018786848.

Lynch, Michael. 2017. ‘STS, Symmetry and Post-Truth’. *Social Studies of Science* 47 (4): 593–99. https://doi.org/10.1177/0306312717720308.

Marres, Noortje. 2018. ‘Why We Can’t Have Our Facts Back’. *Engaging Science, Technology, and Society* 4 (July): 423–43. https://doi.org/10.17351/ests2018.188.

Miceli, Milagros, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. ‘Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices’. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 161–72. FAccT ’21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445880.

Moats, David, and Nick Seaver. 2019. ‘“You Social Scientists Love Mind Games”: Experimenting in the “Divide” between Data Science and Critical Algorithm Studies’. *Big Data & Society* 6 (1): 2053951719833404. https://doi.org/10.1177/2053951719833404.

Monti, Federico, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. 2019. ‘Fake News Detection on Social Media Using Geometric Deep Learning’. *ArXiv:1902.06673 [Cs, Stat]*, February. http://arxiv.org/abs/1902.06673.

Nieminen, Sakari, and Valtteri Sankari. 2021. ‘Checking PolitiFact’s Fact-Checks’. *Journalism Studies* 22 (3): 358–78. https://doi.org/10.1080/1461670X.2021.1873818.

Nyhan, Brendan, and Jason Reifler. 2010. ‘When Corrections Fail: The Persistence of Political Misperceptions’. *Political Behavior* 32 (2): 303–30. https://doi.org/10.1007/s11109-010-9112-2.

Park, Sungkyu, Jaimie Yejean Park, Jeong-han Kang, and Meeyoung Cha. 2021. ‘The Presence of Unexpected Biases in Online Fact-Checking’. *Harvard Kennedy School Misinformation Review*, January. https://doi.org/10.37016/mr-2020-53.

Pérez-Rosas, Verónica, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. ‘Automatic Detection of Fake News’. *ArXiv:1708.07104 [Cs]*, August. http://arxiv.org/abs/1708.07104.

———. 2018. ‘Automatic Detection of Fake News’. In *Proceedings of the 27th International Conference on Computational Linguistics*, 3391–3401. Santa Fe, New Mexico, USA: Association for Computational Linguistics. https://aclanthology.org/C18-1287.

Pollock, Neil, and Robin Williams. 2010. ‘E-Infrastructures: How Do We Know and Understand Them? Strategic Ethnography and the Biography of Artefacts’. *Computer Supported Cooperative Work (CSCW)* 19 (6): 521–56. https://doi.org/10.1007/s10606-010-9129-4.

Prasad, Amit. 2021. ‘Anti-Science Misinformation and Conspiracies: COVID–19, Post-Truth, and Science & Technology Studies (STS)’. *Science, Technology and Society*, April, 09717218211003413. https://doi.org/10.1177/09717218211003413.

Rashkin, Hannah, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. ‘Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking’. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2931–37. Copenhagen, Denmark: Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1317.

Rietdijk, Natascha, and Alfred Archer. 2021. ‘Post-Truth, False Balance and Virtuous Gatekeeping’. In *Virtues, Democracy, and Online Media: Ethical and Epistemic Issues*, edited by Nancy Snow and Maria Silvia Vaccarezza. Routledge.

Sacco, Pier Luigi, Riccardo Gallotti, Federico Pilati, Nicola Castaldo, and Manlio De Domenico. 2021. ‘Emergence of Knowledge Communities and Information Centralization during the COVID-19 Pandemic’. *Social Science & Medicine* 285 (September): 114215. https://doi.org/10.1016/j.socscimed.2021.114215.

Schuurbiers, Daan. 2011. ‘What Happens in the Lab: Applying Midstream Modulation to Enhance Critical Reflection in the Laboratory’. *Science and Engineering Ethics* 17 (4): 769–88. https://doi.org/10.1007/s11948-011-9317-8.

Seifert, Colleen M. 2018. ‘The Distributed Influence of Misinformation.’ *Journal of Applied Research in Memory and Cognition* 6 (4): 397. https://doi.org/10.1016/j.jarmac.2017.09.003.

Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. ‘Fairness and Abstraction in Sociotechnical Systems’. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. FAT* ’19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598.

Shu, Kai, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. ‘FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media’. *Big Data* 8 (3): 171–88. https://doi.org/10.1089/big.2020.0062.

Sloane, Mona, and Emanuel Moss. 2019. ‘AI’s Social Sciences Deficit’. *Nature Machine Intelligence* 1 (8): 330–31. https://doi.org/10.1038/s42256-019-0084-6.

Stewart, Elizabeth. 2021. ‘Detecting Fake News: Two Problems for Content Moderation’. *Philosophy & Technology* 34 (4): 923–40. https://doi.org/10.1007/s13347-021-00442-x.

Suchman, Lucy. 2002. ‘Located Accountabilities in Technology Production’. *Scandinavian Journal of Information Systems* 14 (2). https://aisel.aisnet.org/sjis/vol14/iss2/7.

Tanweer, Anissa, Emily Kalah Gade, P. M. Krafft, and Sarah K. Dreier. 2021. ‘Why the Data Revolution Needs Qualitative Thinking’. *Harvard Data Science Review* 3 (3). https://doi.org/10.1162/99608f92.eee0b0da.

Thacker, Paul D. 2021. ‘The Covid-19 Lab Leak Hypothesis: Did the Media Fall Victim to a Misinformation Campaign?’ *BMJ* 374 (July): n1656. https://doi.org/10.1136/bmj.n1656.

Torabi Asr, Fatemeh, and Maite Taboada. 2019. ‘Big Data and Quality Data for Fake News and Misinformation Detection’. *Big Data & Society* 6 (1): 2053951719843310. https://doi.org/10.1177/2053951719843310.

Uscinski, Joseph E. 2015. ‘The Epistemology of Fact Checking (Is Still Naìve): Rejoinder to Amazeen’. *Critical Review* 27 (2): 243–52. https://doi.org/10.1080/08913811.2015.1055892.

Uscinski, Joseph E., and Ryden W. Butler. 2013. ‘The Epistemology of Fact Checking’. *Critical Review* 25 (2): 162–80. https://doi.org/10.1080/08913811.2013.843872.

Valverde-Albacete, Francisco J., and Carmen Peláez-Moreno. 2014. ‘100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox’. *PLOS ONE* 9 (1): e84217. https://doi.org/10.1371/journal.pone.0084217.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. ‘The Spread of True and False News Online’. *Science* 359 (6380): 1146–51. https://doi.org/10.1126/science.aap9559.

Walter, Nathan, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. 2020. 'Fact-Checking: A Meta-Analysis of What Works and for Whom'. *Political Communication* 37 (3): 350–75. https://doi.org/10.1080/10584609.2019.1668894.

Yearley, Steven. 1999. 'Computer Models and the Public's Understanding of Science: A Case-Study Analysis'. *Social Studies of Science* 29 (6): 845–66.

Zhou, Xinyi, Jindi Wu, and Reza Zafarani. 2020. 'SAFE: Similarity-Aware Multi-Modal FakeNews Detection'. *Advances in Knowledge Discovery and Data Mining* 12085 (April): 354–67. https://doi.org/10.1007/978-3-030-47436-2_27.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for the Future at the New Frontier of Power*. London: Profile Books.

[anonymised references]