REPHRAIN

Protecting citizens online

Ethical, political and epistemic implications of machine learning (mis)information classification: insights from an interdisciplinary collaboration between social and data scientists.

Andrés Domínguez Hernández, University of Bristol

Richard Owen, University of Bristol

Dan Saattrup Nielsen, University of Bristol

Ryan McConville, University of Bristol

July 2023

Routledge
Taylor & Francis Group

Journal of Responsible Innovation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjri20

Ethical, political and epistemic implications of machine learning (mis)information classification: insights from an interdisciplinary collaboration between social and data scientists

Andrés Domínguez Hernández, Richard Owen, Dan Saattrup Nielsen & Ryan McConville

To cite this article: Andrés Domínguez Hernández, Richard Owen, Dan Saattrup Nielsen & Ryan McConville (2023) Ethical, political and epistemic implications of machine learning (mis)information classification: insights from an interdisciplinary collaboration between social and data scientists, Journal of Responsible Innovation, 10:1, 2222514, DOI: 10.1080/23299460.2023.2222514

To link to this article: https://doi.org/10.1080/23299460.2023.2222514

Published online: 07 Jul 2023.

JOURNAL OF RESPONSIBLE INNOVATION
2023, VOL. 10, NO. 1, 2222514
https://doi.org/10.1080/23299460.2023.2222514

Routledge
Taylor & Francis Group

Research Article

OPEN ACCESS Check for updates

Ethical, political and epistemic implications of machine learning (mis)information classification: insights from an interdisciplinary collaboration between social and data scientists

Andrés Domínguez Hernández [a], Richard Owen [b], Dan Saattrup Nielsen [c] and Ryan McConville [c]

[a]Department of Computer Science, University of Bristol, Bristol, United Kingdom; [b]School of Management, University of Bristol, Bristol, United Kingdom; [c]Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom

ABSTRACT

Machine learning (ML) classification models are becoming increasingly popular for tackling the sheer volume and speed of online misinformation. In building these models data scientists need to make assumptions about the legitimacy and authoritativeness of the sources of 'truth' employed 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 public debate and create downstream negative impacts. This article presents findings from a responsible innovation (RI) inflected collaboration between science and technology studies scholars and data scientists. Following an interactive co-ethnographic process, we identify a series of algorithmic contingencies—key moments during ML model development which could lead to different future outcomes, uncertainties and harmful effects. We conclude by offering recommendations on how to address the potential failures of ML tools for combating online misinformation.

ARTICLE HISTORY
Received 6 June 2022
Accepted 5 June 2023

KEYWORDS
And phrases: misinformation; reflexivity; content moderation; fact-checking; machine learning; responsible innovation collaboration

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 content moderators (e.g. Hassan et al. 2017; Monti et al. 2019; Zhou, Wu, and Zafarani 2020). The development of ML tools for content moderation has gained currency particularly

CONTACT Andrés Domínguez Hernández andres.dominguez@bristol.ac.uk Department of Computer Science, University of Bristol, Bristol, United Kingdom

among social media platforms like Facebook, Twitter, TikTok and YouTube not only because of cost and time saving reasons but due to mounting regulatory pressure on platforms to take responsibility for their role in the propagation of harmful content including misinformation (CDEI 2021). 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).

The artificial intelligence (AI) research community has broadly framed the problem as one that can be tackled using ML-enabled classification models. These classify (or 'label'), with varying degrees of accuracy, the category to which a piece of online content belongs – e.g. a claim that is labelled as being 'factually true', 'false' or 'misleading'. The models are trained on large datasets of various modalities (images, video, text or social media metrics) containing human annotated[2] samples of information labelled as being factually correct or false (Torabi Asr and Taboada 2019; Groh et al. 2022). In order to advance the state of the art, researchers strongly emphasise the need for more and higher quality data 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 and videos, or false claims along with labels about their truthfulness produced by researchers or professional fact-checkers.[3]

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; Jaton 2021; Miceli et al. 2021; Selbst et al. 2019; Domínguez Hernández and Galanos 2023). However, scant attention has been given to the *epistemic* assumptions and practices that underlie ML models for misinformation identification and, associated with this, their social, political and commercial entanglements. The prevailing rationale in developing such tools is that if fed with (large and good) enough data they will be able to produce reliable, actionable evaluations of truthfulness, allowing users to tackle the problem of misinformation in an automated, cost-effective manner.

The practice of constructing referential datasets – or 'ground truths' – used for both model training and performance measuring purposes is rooted in assumptions about the credibility and trustworthiness of those data sources. These sources typically include corpora of 'authoritative knowledge' and/or the outputs of professional fact-checking organisations, which are implicitly assumed to be credible and can be used to benchmark what is deemed to be 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.

Our paper addresses this gap. We present findings from a responsible innovation (RI) inflected collaboration between science and technology studies (STS) scholars and data scientists developing ML tools to combat misinformation. This study was undertaken as part of a cross-cutting RI workstream exploring the integration of RI frameworks (Owen et al. 2013; EPSRC 2018) within a large interdisciplinary research Centre aimed at researching and addressing online harms. Our collaboration focused on mobilising the reflexivity and anticipatory dimensions of RI (Owen et al. 2013).

Drawing on insights from social studies of science and using an interactive co-ethnographic approach, our collaboration surfaced a series of *algorithmic contingencies* – key moments during model development which could lead to different future outcomes, uncertainties and even harmful effects. These in turn allowed us to co-develop insights for the responsible development of ML tools for misinformation detection and management. Our frame of contingencies departs from the calculus of fairness or data bias elimination discussed in the literature to date (Selbst et al. 2019; Domínguez Hernández and Galanos 2023). 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, as key goals for responsible innovation.

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 the philosophy of science and social studies of science have compellingly argued, facts are not constructed 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, 103). 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 one socially acceptable means for determining what one might call 'objective knowledge'. Feminist scholars have however warned against the objectivist ideal of 'science as neutral' as it paradoxically ignores 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 per se, but a call to remain cautious about the often loose use of the language of 'neutrality' and 'objectivity' in public discourse (Harding 1995; Lynch 2017). We take this cautionary and critical stance when examining how assumptions about (scientific) knowledge, expertise and facticity might become encoded in ML techniques aimed at labelling, sorting 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 seeking to promote 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.[4]

Modern data-driven fact-checking has increasingly been viewed as being vital to tackle misinformation in the so called 'post truth' era (Carlson 2017). Social media companies, and notably Meta (Facebook's parent company), have partnered with professional fact-checkers to combat the widespread misinformation problem (see Meta 2020). 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 train and refine 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 an inherently 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 are – like scientists – value-laden political actors engaging in epistemic practices: establishing facticity and confronting lies (defined by them) in public discourse (Graves 2017).

Fact-checking services have attracted some criticism over their methodologies due to, for example, accusations of skewed selection of topics and claims, and the use of ambiguous terminology (Uscinski and Butler 2013; Stewart 2021). For instance, fact-checking organisations often use vague or borderline phrases like 'mostly true' or 'mostly false' on the basis that claims are not always verifiable with sufficient certainty as being simply either 'true' or 'false.' 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 is contingent and may change over time, driven by public debate or the emergence 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 (Sayers 2021). This episode shows that while well intentioned, the practice of fact-checking can lead to ambivalences and false positives which could in turn be blindly reproduced and spread by an algorithm.

Not only does the online misinformation ecosystem evolve quickly, but the experiences and manifestations of misinformation differ vastly across cultures, idiosyncrasies, languages and political realities (Prasad 2022; Seifert 2017). 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 misinformation due to the segregating dynamics of online epistemic communities formed around shared politics and identities (Nyhan and Reifler 2010).

In pointing out the challenges involved with establishing the truth we do not seek to undermine the value of expertise, journalism and fact-checking 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 become scaled up and automated.

Methodology: an interdisciplinary responsible innovation collaboration

This article is the result of an interdisciplinary collaboration between data scientists leading a research project (CLARITI) focused on the development of a ML model for tackling online misinformation, and scholars affiliated with the field of science and technology studies (STS) (hereafter the 'research team'). The collaboration was initiated as part of a cross-cutting workstream within a large interdisciplinary research Centre in the UK aimed at researching and addressing online harms.[5] The 'Responsible, Inclusive and Ethical Innovation' workstream aimed to embed Responsible Innovation practices across the Centre, with a focus on interdisciplinary knowledge co-production. This workstream provided a direction, and importantly resources, for our collaboration, which was a voluntary activity engaged by all parties. Our collaboration focused on mobilising the reflexivity (first and second order), anticipatory and responsiveness dimensions of RI, drawing on the AREA framework (Stilgoe, Owen, and Macnaghten 2013; Owen et al. 2013) formalised by the Centres' funder, the UK Engineering and Physical Sciences Research Council.

Our collaboration draws inspiration from 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). It builds upon and contributes to previous efforts to integrate social, ethical and human values considerations into processes of research, design and development (Schuurbiers 2011; Fisher, Mahajan, and Mitcham 2006; van den Hoven 2013)

While ethnography has been the archetypical tool of STS theory and intervention, in this study we explicitly adopted a collaborative configuration of this approach by shifting from an ethnographer/informer arrangement to a joint endeavour between social scientists and data scientists (c.f. Forsythe 1993; Bieler et al. 2021). Collaborative ethnography can be viewed as 'an approach to ethnography that *deliberately* and *explicitly* emphasises collaboration at every point in the ethnographic process, without veiling it – from project conceptualisation, 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 emphasis in original). Our aim with this approach was then not only to enrich the process of critical analysis of technical work through close observation, but to advance collective reflection allowing for the co-development of ethical and responsible ML practices.

Our approach took the form of regular meetings within the research team over a period of approximately 8 months. These were empirically focused on the development of a ML model to detect online misinformation conducted over that same period. The technical project was conducted by a team of data scientists (DSN and RM) and comprised a multimodal ML based study of misinformation on social media and, the development of an ML model for misinformation detection, labelling and management. One of the outcomes of the data science project was a 'misinformation dataset' (Nielsen and McConville 2022) which intends to capture the diverse ways in which misinformation

manifests on social media and which are used to train and validate the ML model. This dataset contains roughly 13,000 claims (of which 95% are labelled as misinformation) from 115 fact-checking organisations 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 newsworthy events (e.g. COVID-19, Israel-Palestine conflicts) appearing on the platform over the course of a decade.

Our collaboration usefully coalesced around the schematisation of the process of building the ML detection model (Figure 1) and the curation of the ground truth datasets used to support this (see Findings). This process of visualisation allowed the social scientists (ADH and RO) to understand and gain literacy in the technical details of the project, the associated development activities and comparable work in the technical literature on automatic misinformation detection.[6] The collaboration combined interpretation, critique, self-critique and calls for action co-produced by the research team. We approached this method iteratively, purposely surfacing technical and epistemic assumptions and practices – more generally in the ML model development literature and more specifically in the project itself. Co-authoring was found to be a productive process to support this, allowing interpretative texts written initially by the social scientists to be checked and expanded by the data scientists.

We acknowledge the limitations of our method in producing generalising claims which are reflective of the research team's specific concerns, experiences, practices within a specific project and a limited subset of works in the literature.

Findings

In the following sections we describe the steps taken in the data science project to construct the automated misinformation detection model. We use this as a tangible way to critically examine epistemic assumptions and practices involved in the development of ML tools for online misinformation detection.

Below we first describe and schematise (Figure 1) the steps in the construction of the model adopted in the project: (1) problem definition, (2) choice of inputs and outputs, (3) curation of ground truth datasets and model training (i.e. 'ground-truthing'), (4) model validation

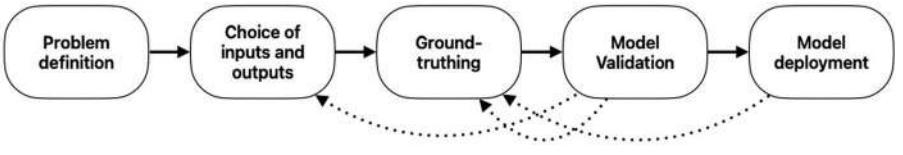

Figure 1. Steps in the construction of a ML misinformation detection model. (1) Problem definition: design of strategy based on hypotheses, definitions and theories about how to identify misinformation. (2) Selection of multimodal inputs and outputs to be included into the classification model. (3) Ground-truthing: the ground truth dataset is used to train the model. A subset of the this is reserved for model validation. (4) Model validation: a subset of the ground truth dataset is used to test the model's performance. Metrics of performance accompany the publication of classification models. (5) Model deployment: model outputs inform online content moderation decisions such as banning, downranking or flagging. The dotted arrows represent feedback loops between steps.

(testing), and finally (5) deployment. The ML model developed here is in essence a *classifier algorithm* that maps inputs (e.g. documents, videos, images) to outputs (e.g. true, false, other) according to examples contained in an annotated ground truth dataset.

We then employ the notion of contingencies to interrogate the entanglements associated with each step in the model development, and the conditions that could alter the actual or claimed utility of a model and lead to potentially deleterious consequences. Harmful impacts could for example include legitimate information being wrongly categorised 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 being a fact (Latour and Woolgar 1986). Thus, our responsible innovation exercise is not aimed at debiasing models 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 instead illustrate the salience of various contingencies so that they are pondered reflexively in the development and auditing of ML tools. In the following sections we examine these contingencies in more detail, describing them first and locating each more specifically in the context of the project's research and development steps (Fig 1). (see sumary in Table 1).

Problem definition

The way misinformation is problematised shapes the strategy used to detect and identify it as well as the 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 into discrete (sometimes binary) categories by a ML model according to sensible evaluations of their truthfulness. These evaluations are influenced and informed by existing empirical studies, theories and hypotheses about what may constitute or signal the presence of misinformation. Categories like misinformation and truth in this sense are defined a-priori by the researchers and then translated (formalised) into a ML model's variables (see below) following different data-driven strategies which could leverage e.g. a corpus of knowledge, writing style or patterns of online content propagation (Gradoń et al. 2021). For example, 'fake news' – a popularised 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 posts). 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.

Table 1. A summary of cautions and contingencies in the development of ML misinformation classification models and their implications.

Steps	Cautions and contingencies	Implications
Problem definition	Researchers use different theories and definitions of misinformation. Problematisation relies on previous empirical studies and experiments. Researchers often use hypotheses of how misinformation manifests online.	Misinformation and truth are conceptualised as classifiable based on salient features and prior knowledge. Definitions and provisional hypotheses translate into intractable variables and norms embedded in a model.
Choice of inputs and outputs	The identification of inputs (e.g. text, image, a combination of both) and outputs ('true', 'false', 'misleading') is informed by a combination of technical and cost factors for accessing data.	Opportunity and feasibility constrain the choices of variables and how the problem is abstracted into a mathematical formalism. Proxy variables may be used due to reduced cost (e.g. humour as a signal of misinformation).
Ground-truthing	Ground truth datasets are temporally contingent and have diminishing returns vis à vis new events. Researchers might standardise labels to reduce complexity. Data annotation and labelling are susceptible to errors, partiality and uncertainty (e.g. factcheckers disagreeing on the verdict of a claim). Using fact-checks to curate datasets could overrepresent debunked claims over truthful claims.	Difficulty to check for claims relating to new events. Inconsistency among data sources leading to loss of nuance or amplification of errors. Models are better at recognising debunked claims and less so at checking truths. Selection biases influence what types of information are worthy of checking and which narratives are prioritised over others.
Model Validation	Performance evaluations are only indicative of novelty and progress among competing models. Benchmarks of performance against other models could be used across content domains.	Performance metrics influence moderation decisions such as censoring, downranking, flagging pieces of information. Model high performance can be a misleading metric of a model's utility in particular topics and cultural contexts.
Model deployment	Moderation decisions can be informed on metrics of ML model performance. A logic of scaling up could lead to blanket decisions across cultures, locations, idiosyncrasies, laws. Utilitarian approach: some error or harm (false positives and false negatives) is acceptable.	High performing models leading to less human input. Using high performance models as a solution might underplay the responsibilities of platforms. Models could have uneven impacts across different contexts. Burden of algorithmic failures and harm shifted away from platforms e.g. appealing processes

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 comparison with those exhibited by legitimate, 'truthful' information. For instance, several 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 speed of consumption of falsehoods that could be used to infer the presence of misinformation (Monti et al. 2019).

In the CLARITI project, multiple hypotheses underpinned the construction of the model which drew 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 (Cui and Lee 2020; Li et al. 2020b); and (6) a classifier trained on data which is monolingual or monotopical will not generalise to new languages and events (Han, Karunasekera, and Leckie 2020).

In sum, definitions, prior evidence and hypotheses (even if they are provisional) all inform further steps in the construction of the model and thereby become nested in the pipeline of development and which are only auditable if explained sufficiently in a technical paper or documentation.

Choice of inputs and outputs

Once the problem is defined, it is formalised (or abstracted) into a mathematical function with inputs and outputs so that it can then be operationalised computationally. The choice of data inputs and outputs (also called target variables) for a ML model not only reflects the researcher's framing of the problem but also technical feasibility in terms of what kind of data can be reliably and economically acquired and used at scale. The formalisation of the problem and the availability of training data are in this sense mutually constitutive; or as Selbst et al. (2019, 60) put it, 'abstraction choices often occur implicitly, as accidents of opportunity and access to data.' For instance, while most algorithmic techniques to date have used text as an input variable, misinformation is increasingly 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 interest in working with multimodal data, which was indeed also a feature of the CLARITI project.

Lack of data is frequently mentioned in the literature as one of the biggest obstacles to 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. Meta) while others make this somewhat easier (e.g. Twitter where academic access is 10 million tweets per month). On the other hand, in order to use supervised ML models, online content needs to be labelled to specific output categories e.g. as 'true', 'false', 'misleading', and so on (see Ground-truthing below).

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 sites 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, they are susceptible to making spurious associations and are therefore prone to failure. 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 humour (Pérez-Rosas et al. 2018).

Opportunity, resources and feasibility were key considerations within the CLARITI project. The strategy was to use a *feature-rich* set of inputs 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 outputs were defined as 'factual' and 'misinformation' and anything outside this dichotomy was categorised as 'other'. The model ultimately relied on annotated, multimodal data (i.e. texts and images) obtained from the Google Fact Check Explorer[7] as well as from Twitter which provides access to data about user engagements with news. This was motivated by the ready availability of the data for research purposes, but also because the widespread use of these sources in the automatic misinformation detection research literature would make it easier to test how the current effort would compare against existing models.

Ground-truthing

Having defined the strategy and the choice of inputs and outputs, the next step is to train the ML model using a ground truth dataset as a referent of factual and non-factual information. This process of *ground-truthing* is not a trivial act but requires the ML developer to make normative choices as to what is an acceptable source of legitimate information (Jaton 2021). 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 variously used to build labelled datasets of categorised (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;[8] and journalists might (unintendedly or not) publish data in a way that is skewed and misleading (Lewis and Westlund 2015). For illustration, here we enumerate some of the conditionalities and contingencies 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 could shift dramatically over a short period of time. These shifts – which may call for rectification – are not always adequately and consistently addressed by fact-checkers such that they can be taken on board in updating an ML model.[9] If left unchecked, ambivalences in the training data could lead to the amplification of objectionable results and potentially harmful false positives. Yet determining the impact of ambivalences, let alone checking and correcting them, can be a challenging and costly endeavour.

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 may not be blatant or obvious, but subtle and nuanced. Different organisations 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 interpret, standardise or develop new labels from existing data. For instance, to tackle the issue with inconsistent labels in the CLARITI project, 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' – 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 standardised even if this introduces new ambiguities and loss of nuance. For example, 'half true' was categorised as 'misinformation'. To mitigate these ambiguities, a decision was made in the project to only include claims whose verdicts from the fact-checking organisations 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.[10] 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 address 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 the case in the CLARITI project 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 articles from 'trusted sources', as this would both introduce more bias as well as potentially *polluting* data from a different data distribution. In other words, in attempting to balance the data, ML models could end up distinguishing 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 organisations. For instance, a comparative study of two major fact-checking organisations 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, cultural norms and languages are prioritised over others (Duarte, Llanso, and Loup 2018).

Model validation

The merit and utility of a classification model is judged by its ability to accurately 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 dataset. For example, in the case of models using datasets with labels provided by fact-checkers, 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*[11]) 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 decontextualised; 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 decontextualisation 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, 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 CLARITI project's ground truth 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 for more consistent evaluations, albeit substantially harder. Further, as mentioned above, 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*[12] because 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. (2018) models with over 70% 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 misinformation detection systems. They can either configure hybrid decision-support systems (e.g. ML-assisted fact-checking) or operate as standalone, automated moderation systems with no human input. In the case of Meta, content initially 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, which leads platforms increasingly to rely on automation.[13]

Depending on the platform's moderation policy, automated detection systems 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') interests and definitions of what counts as being acceptable. Moreover, there is widespread evidence that major social media platforms have facilitated the formation of online 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, platform-wide corrective actions may not only have disparate effects when used across different groups, languages and cultural contexts but paradoxically pose the risk of reinforcing false beliefs particularly amid those online epistemic communities where the circulation of falsehoods or conspiracy theories is more prevalent (Madraki et al. 2021; 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 who can emulate the style of truthful sources or translate 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 optimisation. 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 emphasises the role of individual consumers and producers 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 maximising the time users spend on their platforms in order to generate advertising revenue. This is achieved through opaque algorithms of personalisation 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). One could argue that the commercial incentive of platforms to maximise engagement is thus at odds with the goal of meaningfully tackling the spread of misinformation and any type of content that could lead to downstream harms.

Furthermore, there is a risk that automation is positioned as the only solution to the spread of harmful content as it ensures efficient business as usual. While we recognise that machine learning models can have promising benefits to deal with the scale of misinformation, their development should not be viewed as a final solution to the problem nor should they foreclose the broader debate around platform regulation and oversight.

Embedding responsible practices in algorithmic content moderation

Interdisciplinary interactions and collaborations between data scientists and social scientists (from health and biological sciences to computing and robotics) are not new. However, the initiation of such collaborations aimed at fostering reflexive and responsive development of ML models *in hand with the technical development* is much rarer.

Our collaboration surfaces a series of contingent epistemic practices, institutional commitments and socially constructed assessments of facticity that suggest 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 emphasise that misinformation detection algorithms are highly time sensitive: models using historic data may quickly become obsolete especially in relation to new events. This makes it difficult for ML models to maintain their utility unless they are routinely re-trained with up-to-date information and attuned to 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 minimising any harmful effects. There is no straightforward procedure to establish what the *right* outcomes might look like given that desired outcomes for platforms and regulators, downstream harmful effects and social preferences toward misinformation might be in conflict. For instance, while some might be in favour of reducing the volume of misinformation online by maximising 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 socio-technical, cultural, legal, political and epistemic considerations. We suggest developers should endeavour to account for algorithmic contingencies and sufficiently document 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 problem formulation, choice of variables, ground-truthing, and model 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 that emerged during our collaboration aimed at engendering responsible innovation in this field.

(a) Reflexivity beyond datasets

Principled and defensible criteria such as relevance, authoritativeness, data structure and timelines of truth assessments all provide a strong foundation for the curation and use of ground truth datasets. There are already 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 underpin and configure the development and use of ML models.

A recent study by Birhane et al. (2022) showed that highly cited ML research has typically ascribed to values of performance, efficiency and novelty over considerations of social needs, harms and limitations and that researchers frequently 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 design choices, assumptions and methodological 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 dimension of responsible innovation (Stilgoe, Owen, and Macnaghten 2013), and a recurring theme in many papers published in the current journal (e.g. Foley, Sylvain, and Foster 2022) – 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 (Owen et al. 2021). 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 should seek to avoid shifting the blame on the data and external sources, acknowledge partiality (as opposed to deceptive attempts 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 raise the need for transparency reports that are accompanied by reflexive disclaimers about developers' methodological choices, problem statements, institutional affiliations and sources of funding influencing data collection and model construction.

(b) 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 cultural norms, laws and regulations. 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 might decrease in size over time – as the training and test data changes, the model performance will change. This would allow for some form of dynamic and adaptive environment, where published models, their results and the specific role of human moderation are continuously re-evaluated and verdict changes are reflected as appropriate. Community benchmarks based on location, language and domain-specificity are one way to encourage this. These evaluations should be consistent so that models are assessed with attention to topicality, timing, context, language or different modalities of content 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.

Calls for algorithmic audits and due diligence are now gaining traction in the AI ethics debate, particularly in areas of application which have well-documented risks and harms (Brown, Davidovic, and Hasan 2021). Less attention however has been given to applications that are positioned as socially beneficial and pressing such as content moderation. In the case of misinformation in particular, further research is needed around how conditioning factors such as fact-checkers' political leanings, domain specialisms but also local laws and specific political demands around truth and freedom of expression could be factored into the mix of quantitative or qualitative algorithmic evaluations.

(c) Accounting for and communicating uncertainty

Acknowledging and responding to uncertainty is a key aspect of responsible innovation (Stilgoe, Owen, and Macnaghten 2013). This is particularly critical in the case of machine

learning algorithms, where uncertainty is an inherent property. ML researchers have tried to understand and measure uncertainty in different ways (Hüllermeier and Waegeman 2021; Bhatt et al. 2021). Qualitative, also called 'epistemic', uncertainty extends to the lack of knowledge about the outputs of a model or ignorance on the part of the decision maker about the innerworkings of a ML model. This type of uncertainty can be a reflection of several factors that are baked into a model such as subconscious biases, inaccuracies or gaps in the data, as well as discretionary forms of data reduction or standardisation of labels that are sometimes carried out by developers. For instance, 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 uncertainty such as casting doubt on legitimate information, doing away with important nuances in language, or leaving subtle misleading claims unaddressed.

ML model construction is also invariably impacted by randomness, or what is known by ML researchers as 'aleatoric' uncertainty. For instance, 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 academic research (thus two people searching for 'coronavirus' using that API may collect different results and thus create a different dataset). 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.

Measuring and reporting the various forms of uncertainty in a model can be crucial both for aiding human intervention and for increasing the overall transparency of the system. While it might be challenging to comprehensively address all uncertainties, some measures can be taken to help communicate it to others. Any quantitative measures of uncertainty should be published in tandem with other metrics as part of the model evaluation. This can not only help avoid overreliance on algorithms and minimise ambiguous outcomes by helping human moderators but also help the pondering of alternative interventions such as adding context to ambiguous content or links to contrasting news. In addition, the data collection systems themselves should be made public and available. Even if not fully reproducible (due to the dynamic nature of social media platforms), developers should provide complete auditable documentation on the data collection process. This approach was indeed considered as a way forward in the CLARITI project where the data collection system that was used was made available on GitHub (Nielsen and McConville 2022) so others can see and execute the exact code used to build the dataset, and thus all decisions that were made.

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 identified contingencies and insights to support 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 promising developments, these tools are highly contingent on the epistemic status of their

ground truths, and the assumptions, choices and definitions underpinning tool 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 being 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 violent content. This marks paths of future inquiry where our analytical approach could be used to interrogate the epistemologies and forces driving other algorithmic systems.

Our 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 reflect on the methodological learning from the collaboration in terms of practicing RI in a companion paper to the present study (Domínguez Hernández and Owen forthcoming). Therein we highlight (a) the value of collective self-critique as a key step in moving to action within responsible innovation collaborations; (b) the importance of gaining technical literacy for the innovation under consideration, facilitating a move from initial 'strategic vagueness' in the early stages of the collaboration to a more technically informed critical synthesis with the identification of tangible insights and recommendations, and (c) employing co-authorship as a practice that productively supports knowledge co-creation across disciplines.

We hope our contribution sparks further exploration of how data scientists and social scientists can work together so as to break with longstanding, yet 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). In considering the wealth of literature on RI, STS, critical algorithm studies and AI ethics, we also recognise that previous learning cannot be taken as being readily intelligible and applicable to those expected to apply said learning and therefore that in-situ inquiries are imperative. We thus approached this study in an open-ended and co-productive fashion, conducting iterative cycles of observation, interpretation, validation and calibration among the research team. We call for further experimentation of this type that supports the embedding of RI within data and computational sciences and widens the audience and genres of contributions in RI.

Notes

1. While several terms related to misinformation (e.g., disinformation and fake news) are used throughout this paper, 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.
2. Machine learning algorithms allow computers to 'learn' based on examples derived from data. This process demands human labour—in what is known as supervised ML—to develop training and testing datasets containing pieces of information typically annotated by humans.
3. A non-exhaustive list of datasets is found on D'Ulizia et al. (2021)

4. A database of global fact-checking websites has identified more than 300 (Reporter's Lab 2022)
5. The National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online REPHRAIN.
6. We note that some of the technical papers reviewed here are at the time of writing still in preprint form, which is instructive of the faster pace of AI development compared with the academic peer review and publication cycles.
7. This service aggregates claims which have been fact-checked by eligible news organisations. To be included in Google's fact check tool, news organisations need to comply with Google's standards for publishers and content policies (Google 2023).
8. An illustration of this is what Rietdijk and Archer (2021) problematise 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.
9. For example, PolitiFact, a well-known fact-checker organisation, decided to archive their original assessment on the COVID lab leak controversy by removing it from their database and revising their assessment as 'widely disputed' (see PolitiFact 2021)
10. Some fact-checking organisations focus exclusively on false and misleading claims (e.g., factcheck.org)
11. 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.
12. 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).
13. This recursive feedback loop involving algorithms and humans influencing one another introduces yet further contingencies, however we do not have space here to explore these in detail.
14. As 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)

Disclosure statement

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

Funding

This research 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.

Notes on contributors

Andrés Domínguez Hernández is Senior Research Associate at the University of Bristol, United Kingdom. He has a background in science and technology studies, technology and innovation management, and electronics engineering. He has a longstanding interest in interdisciplinary collaborations across engineering, computer science, design and the social sciences. He has taken part in international projects investigating responsibility, justice and ethics with emerging digital technologies. Prior to academia, he led innovation and technology transfer policy and the implementation of large-scale telecommunications infrastructure.

Richard Owen is a Professor of Innovation Management in the School of Management, Faculty of Social Sciences, University of Bristol, U.K. He is interested in the power of innovation and techno-visionary science to create futures in profound and uncertain ways, how we can engage as a society with those futures and how we can take responsibility for them. He is interested in the politics, risks, ethics and governance of innovation and new technologies in society. His research sits at the intersection of innovation governance and science and technology studies as a critical, interdisciplinary scholar.

Dan Saattrup Nielsen is a postdoctoral researcher in machine learning at the University of Bristol, where he also was awarded his PhD in Mathematics. Previously, he has been working with graph machine learning for fraud detection at the Danish Business Authority. His research interests include graph machine learning and natural language processing for low-resource languages

Ryan McConville was appointed a Lecturer in Data Science, Machine Learning and AI within the Intelligent Systems Laboratory and Department of Engineering Mathematics at the University of Bristol in September 2019. He gained his PhD working with the Centre for Secure Information Technologies (CSIT) at Queen's University Belfast in 2017 where he researched large scale unsupervised machine learning for complex data. He has worked with inter-disciplinary academic and industrial partners on numerous projects, including large-scale fraud detection and large-scale pervasive personal behaviour analysis for clinical decision support. His research interests lie around unsupervised machine learning, deep learning on multimodal and complex data with applications to social network analysis, recommender systems, healthcare and cybersecurity.

ORCID

Andrés Domínguez Hernández http://orcid.org/0000-0001-7492-7923
Richard Owen http://orcid.org/0000-0002-1767-3901
Dan Saattrup Nielsen http://orcid.org/0000-0001-9227-1470
Ryan McConville http://orcid.org/0000-0002-7708-3110

References

Torabi Asr, Fatemeh, and Maite Taboada. 2019. "Big Data and Quality Data for Fake News and Misinformation Detection." *Big Data & Society* 6 (1): 205395171984331. https://doi.org/10.1177/2053951719843310.

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.

Bhatt, Umang, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, et al. 2021. "Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–413. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3461702.3462571

Bieler, Patrick, Milena D. Bister, Janine Hauer, Martina Klausner, Jörg Niewöhner, Christine Schmid, and Peter. Sebastian von. 2021. "Distributing Reflexivity Through Co-Laborative

Ethnography." *Journal of Contemporary Ethnography* 50 (1): 77–98. https://doi.org/10.1177/0891241620968271.

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 (Lecture Notes in Computer Science)*, edited by Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri, 405–415. 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. 2022. "The Values Encoded in Machine Learning Research." In *In 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, 173–184. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533083

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): 285–298. https://doi.org/10.1080/09537320600777002.

Brown, Shea, Jovana Davidovic, and Ali Hasan. 2021. "The Algorithm Audit: Scoring the Algorithms That Score Us." *Big Data & Society* 8 (1): 205395172098386. https://doi.org/10.1177/2053951720983865.

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'. Centre for Data Ethics and Innovation. 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* 9), https://doi.org/10.1073/pnas.2023301118.

Collins, H. M., and Robert Evans. 2002. "The Third Wave of Science Studies." *Social Studies of Science* 32 (2): 235–296. https://doi.org/10.1177/0306312702032002003.

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.

Domínguez Hernández, Andrés, and Richard Owen. forthcoming. '"We Have Opened a Can of Worms": Reconciling Critique and Design of Algorithmic Systems Through Co-Ethnography'. Journal of Responsible Innovation, Critique in, of, and for Responsible Innovation,, no. Critique in, of, and for Responsible Innovation.

Domínguez Hernández, Andrés, and Vassilis Galanos. 2023. 'A Toolkit of Dilemmas: Beyond Debiasing and Fairness Formulas for Responsible AI/ML'. In IEEE International Symposium on Technology and Society 2022 (ISTAS22). IEEE. https://doi.org/10.48550/arXiv.2303.01930

Dou, Yingtong, Kai Shu, Congying Xia, Philip S. Yu, and Lichao Sun. 2021. "User Preference-Aware Fake News Detection." In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2051–2055. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3404835.3462990

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, PMLR*, 106–106. https://proceedings.mlr.press/v81/duarte18a.html

D'Ulizia, Arianna, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. 2021. 'Repository of Fake News Detection Datasets'. 4TU.ResearchData. doi:10.4121/14151755.v1.

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, IMC '21*, 444–463. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3487552.3487859

EPSRC. 2018. 'Framework for Responsible Innovation'. 2018. https://www.ukri.org/about-us/epsrc/our-policies-and-standards/framework-for-responsible-innovation/.

Fisher, Erik, Roop L. Mahajan, and Carl Mitcham. 2006. "Midstream Modulation of Technology: Governance from Within." *Bulletin of Science, Technology & Society* 26 (6): 485–496. https://doi.org/10.1177/0270467606295402.

Foley, Rider W., Olivier Sylvain, and Sheila Foster. 2022. "Innovation and Equality: An Approach to Constructing a Community Governed Network Commons." *Journal of Responsible Innovation* 9 (1): 49–73. https://doi.org/10.1080/23299460.2022.2043681.

Forsythe, Diana E. 1993. "Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence." *Social Studies of Science* 23 (3): 445–477. https://doi.org/10.1177/030631279302300302.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92. https://doi.org/10.1145/3458723.

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, FAT* '20*, 325–336. 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): 205395172094323. https://doi.org/10.1177/2053951720943234.

Google. 2023. 'Google News Policies - Publisher Center Help'. 2023. https://support.google.com/news/publisher-center/answer/6204050.

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): 205395171989794. 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): 205395172110138. 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): 201–213. https://doi.org/10.1016/j.eswa.2019.03.036.

Graves, Lucas. 2017. "Anatomy of a Fact Check: Objective Practice and the Contested Epistemology of Fact Checking." *Communication, Culture & Critique* 10 (3): 518–537. https://doi.org/10.1111/cccr.12163.

Groh, Matthew, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. "Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds." *Proceedings of the National Academy of Sciences* 119 (1): e2110013119. https://doi.org/10.1073/pnas.2110013119.

Hall, Kita. 1999. "Performativity." *Journal of Linguistic Anthropology* 9 (1/2): 184–187. https://doi.org/10.1525/jlin.1999.9.1-2.184.

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. "Simians, Cyborgs, and Women." *In Simians, Cyborgs, and Women*, https://doi.org/10.4324/9780203873106.

Harding, Sandra. 1995. "?Strong Objectivity?: A Response to the New Objectivity Question." *Synthese* 104 (3): 331–349. https://doi.org/10.1007/BF01064504.

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 KDD '17*, 1803–1812. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3097983.3098131

Hoven, Jeroen van den. 2013. "Value Sensitive Design and Responsible Innovation." In *Responsible Innovation*, 75–83. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118551424.ch4

Hüllermeier, Eyke, and Willem Waegeman. 2021. "Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods." *Machine Learning* 110 (3): 457–506. https://doi.org/10.1007/s10994-021-05946-3.

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): 205395172110135. 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, MM '17*, 795–816. 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 (Vol. 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." *Digital Journalism* 3 (3): 447–466. https://doi.org/10.1080/21670811.2014.976418.

Li, Yichuan, Bohan Jiang, Kai Shu, and Huan Liu. 2020. "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.

Li, Yichuan, Bohan Jiang, Kai Shu, and Huan Liu. 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): 205316801878684. https://doi.org/10.1177/2053168018786848.

Lynch, Michael. 2017. "STS, Symmetry and Post-Truth." *Social Studies of Science* 47 (4): 593–599. https://doi.org/10.1177/0306312717720308.

Madraki, Golshan, Isabella Grasso, Jacqueline M. Otala, Yu Liu, and Jeanna Matthews. 2021. "Characterizing and Comparing COVID-19 Misinformation Across Languages, Countries and Platforms." In *Companion Proceedings of the Web Conference 2021, 213–23. WWW '21*. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442442.3452304

Marres, Noortje. 2018. "Why We Can't Have Our Facts Back." *Engaging Science, Technology, and Society* 4 (July): 423–443. https://doi.org/10.17351/ests2018.188.

Meta. 2020. 'Here's How We're Using AI to Help Detect Misinformation'. 19 November 2020. https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/.

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 FAccT '21*, 161–172. 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): 205395171983340. 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.

Nielsen, Dan S., and Ryan McConville. 2022. "MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset." In *Proceedings of the 45th*

International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR '22, 3141–3153. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3477495.3531744

Nieminen, Sakari, and Valtteri Sankari. 2021. "Checking PolitiFact's Fact-Checks." *Journalism Studies* 22 (3): 358–378. 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): 303–330. https://doi.org/10.1007/s11109-010-9112-2.

Owen, Richard, Jack Stilgoe, Phil Macnaghten, Mike Gorman, Erik Fisher, and Dave Guston. 2013. "A Framework for Responsible Innovation." In *Responsible Innovation*, 27–50. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118551424.ch2

Owen, Richard, Mario Pansera, Phil Macnaghten, and Sally Randles. 2021. "Organisational Institutionalisation of Responsible Innovation." *Research Policy* 50 (1): 104132. https://doi.org/10.1016/j.respol.2020.104132.

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. 2018. "Automatic Detection of Fake News." In *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.

PolitiFact. 2021. 'Archived Fact-Check: Tucker Carlson Guest Airs Debunked Conspiracy Theory That COVID-19 Was Created in a Lab', 2021. https://www.politifact.com/li-meng-yan-fact-check/.

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): 521–556. https://doi.org/10.1007/s10606-010-9129-4.

Prasad, Amit. 2022. "Anti-Science Misinformation and Conspiracies: COVID–19, Post-Truth, and Science & Technology Studies (STS)." *Science, Technology and Society* 88 (April)), 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 *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2931–2937. Copenhagen, Denmark: Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1317

Reporter's Lab. 2022. 'Database of Global Fact-Checking Sites'. Duke Reporters' Lab. 2022. https://reporterslab.org/fact-checking/.

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.

Sayers, Freddie. 2021. 'Facebook Censors Award-Winning Journalist for Criticising the WHO'. UnHerd, 11 February 2021. https://unherd.com/thepost/facebook-censors-award-winning-journalist-for-criticising-the-who/.

Schuurbiers, Daan. 2011. "What Happens in the Lab: Applying Midstream Modulation to Enhance Critical Reflection in the Laboratory." *Science and Engineering Ethics* 17 (4): 769–788. https://doi.org/10.1007/s11948-011-9317-8.

Seifert, Colleen M. 2017. "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 *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): 171–188. https://doi.org/10.1089/big.2020.0062.

Sloane, Mona, and Emanuel Moss. 2019. "AI's Social Sciences Deficit." *Nature Machine Intelligence* 1 (8): 330–331. https://doi.org/10.1038/s42256-019-0084-6.

Stewart, Elizabeth. 2021. "Detecting Fake News: Two Problems for Content Moderation." *Philosophy & Technology* 34 (4): 923–940. https://doi.org/10.1007/s13347-021-00442-x.

Stilgoe, Jack, Richard Owen, and Phil Macnaghten. 2013. "Developing a Framework for Responsible Innovation." *Research Policy* 42 (9): 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008.

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 Methods." *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.

Uscinski, Joseph E. 2015. "The Epistemology of Fact Checking (Is Still Naïve): Rejoinder to Amazeen." *Critical Review* 27 (2): 243–252. 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): 162–180. 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): 1146–1151. 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): 350–375. https://doi.org/10.1080/10584609.2019.1668894.

Yearley, Steven. 1999. "Computer Models and the Public's Understanding of Science." *Social Studies of Science* 29 (6): 845–866. https://doi.org/10.1177/030631299029006002.

Zhou, Xinyi, Jindi Wu, and Reza Zafarani. 2020. "Lecture Notes in Computer Science." *Advances in Knowledge Discovery and Data Mining* 12085 (April): 354–367. 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.