Behaviour & Information Technology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tbit20

Ethical and safety considerations in automated fake news detection

Benjamin D. Horne, Dorit Nevo & Susan L. Smith

To cite this article: Benjamin D. Horne, Dorit Nevo & Susan L. Smith (04 Dec 2023): Ethical and safety considerations in automated fake news detection, Behaviour & Information Technology, DOI: 10.1080/0144929X.2023.2285949

To link to this article: https://doi.org/10.1080/0144929X.2023.2285949

Published online: 04 Dec 2023.

BEHAVIOUR & INFORMATION TECHNOLOGY
https://doi.org/10.1080/0144929X.2023.2285949

Taylor & Francis
Taylor & Francis Group

Ethical and safety considerations in automated fake news detection

Benjamin D. Horne [a,b], Dorit Nevo[c] and Susan L. Smith[d]

[a]School of Information Sciences, University of Tennessee, Knoxville, TN, USA; [b]Data Science and Engineering, The Bredesen Center, University of Tennessee, Knoxville, TN, USA; [c]Lally School of Management, Rensselaer Polytechnic Institute, Troy, NY, USA; [d]Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY, USA

ABSTRACT

This paper highlights ethical issues in automated fake news detection and calls for caution when deploying tools to automatically detect mis/disinformation in real-life settings. We argue that the potential harm to information consumers caused by an automated tool making a mistake requires us to better understand the mistakes that can be made. We implement three proposed detection models from the literature that were trained on over 381,000 news articles published over six months. We test each of these models using a test dataset constructed from over 140,000 news articles published a month after each model's training data. Articles in the test dataset could come from any outlet, no matter if that outlet was labelled during training or never used during training. We used these data to explore and understand two specific problems with algorithmic fake news detection, namely Bias and Generalisability. These problems arise from the models' training, design, and the inherent unpredictability of news content. Based on our analysis, we discuss the importance of understanding how ground truth is determined, how operationalisation may perpetuate bias, and how the simplification of models may impact the validity of predictions. We offer avenues for future research.

ARTICLE HISTORY
Received 2 May 2023
Accepted 15 November 2023

KEYWORDS
Automated fake news detection; content moderation; algorithmic bias; ground truth; ethics; machine learning

1. Introduction

Fake news is a popular term, broadly used to refer to news articles that are 'intentionally and verifiably false, and could mislead readers' (Allcott and Gentzkow 2017). While the intentionality of fake news is an important part of its definition, it is not uncommon to see it applied to unintentionally misleading information. Consequently, we refer to fake news as encompassing two related terms, namely, misinformation and disinformation. While both refer to untrue information, the former is simply wrong or misleading information, whereas the latter is known to be deliberately false (Lazer et al. 2018; Stahl 2006). Regardless of the original intention of the content creator, both misinformation and disinformation lead to similar outcomes, which is the spread of untrue, partially untrue, or decontextualised information. Further, research on the spread of fake news (on Twitter) showed that 'falsehood diffused significantly farther, faster, deeper, and more broadly than the truth' (Vosoughi, Roy, and Aral 2018).

Due to the overwhelming scale of information online, there have been many proposed technical solutions to combating fake news, the vast majority of which have been developed as detection systems. Proposed detection systems can filter out or automatically place warning labels on news that is of low veracity. These solutions range widely in terms of technical methods used, including various types of Machine Learning (ML) models using the content in news articles and claims (Barrón-Cedeno et al. 2019; Baly et al. 2018; Horne and Adali 2017; Horne, Nørregaard, and Adali 2019), network-based models for news outlets and social media accounts (Gruppi et al. 2022b; Shu, Wang, and Liu 2018), knowledge-graph models for fact-checking (Ciampaglia et al. 2015), and even opaque Large Language Models (LLMs), like ChatGPT, for outlet-level credibility predictions (Yang and Menczer 2023). Across these technical methods, an extraordinary number of classifiers have been proposed, many with unique designs. As an example of this volume, between 2016 and 2023 there were approximately 14,500 papers indexed by Google Scholar that use the phrase 'fake news detection'[1] and 210,000 papers that used the phrase 'fake news'.

The extraordinary number of fake news classifiers proposed demonstrates a problem that is at the heart of a broader critique of the ML community: *leaderboard culture, 'SOTA-chasing'* (Rodriguez et al. 2021) or

CONTACT Benjamin D. Horne bhorne6@utk.edu School of Information Sciences, University of Tennessee, 1331 Circle Park Drive, Knoxville, TN 37996-0332, USA; Data Science and Engineering, The Bredesen Center, University of Tennessee, 1331 Circle Park Drive, Knoxville, TN 37996-0332, USA

'*leaderboardism*' (Ethayarajh and Jurafsky 2020; Hutchinson et al. 2022). Across a variety of ML tasks, including fake news detection, research has been built around optimising models using benchmark datasets. Models are ranked by performance metrics such as accuracy, precision, and recall on those benchmark datasets – creating a *leaderboard*. Subsequently, many papers are then built on the best performing model – the *state-of-the-art (SOTA)*. This format has certainly helped push technical advances in a variety of areas, but the format often sacrifices empirical rigour and neglects the deployment context of the models (Ethayarajh and Jurafsky 2020; Hutchinson et al. 2022; Sculley et al. 2018), which ultimately raises some ethical concerns. Specifically, while many of the proposed automated approaches for fake news detection have shown high accuracy in lab settings, they may be overfitting to specific benchmark data. The construction of benchmark datasets heavily relies on the assessment of *ground truth*.

Ground truth refers to a set of data labels used to train and evaluate ML models. How these labels are generated is highly dependent on the specific ML application. Often, these labels are generated through human expertise. In some cases, the labels are generated with certainty. For example, if we built a model to predict if a widget in a manufacturing line is defective, we know what a *good* widget looks like, and we likely have a clear idea of what *bad* widgets could look like. Our real-world input distribution is well-understood, its scope is limited, and the labels are rigid. However, in applications like fake news detection, experts experience high degrees of uncertainty when labelling data, which may impact the downstream predictions from the ML tool (Bozarth, Saraf, and Budak 2020; Lebovitz, Levina, and Lifshitz-Assaf 2021).

Taken together, the leaderboard culture and the underlying ground truth uncertainty in fake news detection can perpetuate biases from the data, and therefore perpetuate harm and diminish internal validity (Birhane et al. 2022; Bowman and Dahl 2021; Carter et al. 2021; Koch et al. 2021; Liao et al. 2021; Rodriguez et al. 2021; Zhang, Harman, et al. 2020). Further, evaluating models in a simplified context and applying those same models to an uncertain deployment setting can limit external validity and generalisability. By relying on narrowly defined ground truth, performance metrics, and a small number of benchmark datasets, researchers have focused on evaluating techniques in limited, learner-specific, decontextualised settings. If these proposed tools are going to be deployed in real systems, a critical discussion of *what could go wrong* needs to be had. This presents a critical gap in the literature, which we address in this paper using an ethics lens.

While there is a growing body of work criticising the reliance on ground-truth training and decontextualised deployment in public employment services (Dahlin 2021; Sztandar-Sztanderska and Zielenska 2018), judicial bail decisions (Lakkaraju et al. 2017), and medical diagnoses (Lebovitz, Levina, and Lifshitz-Assaf 2021), comparatively little work has critiqued proposed systems in automated content moderation. Some of the critiques of AI may be applicable across domains, for example, the reliance on black-box models (Asatiani et al. 2020; Beltramin, Lamas, and Bousquet 2022; Dahlin 2021; Riley 2019; Wadden 2022), however, identifying issues within the specific deployment setting of an automated system is needed. Of the little work that has occurred in this area, important issues in model training and evaluation have been shown. Specifically, Bozarth, Saraf, and Budak (2020) showed that both model performance and bias can vary drastically based on the choice of training data sets and ground truth labels. Work has also explored the impact of algorithmic transparency on human interactions with automated content moderation (Epstein et al. 2022, Horne, Nevo, et al. 2019), but these works focused on controlled experiments rather than real-life deployment.

We add to this early literature by having a critical discussion of two ethical issues in deploying automated fake news classifiers. First, *algorithmic bias* may be present and systematically skew judgement of news reliability towards, or away from, certain articles, certain topics, or certain sources. While algorithmic bias has been broadly discussed in human-centered contexts (a famous example is that of the recidivism algorithm shown to be racially biased) it has not been explored much in the content moderation context. Second, in traditional ML, tools are tested on data that comes from the same distribution as the training examples – samples that are independently and identically distributed, often called the *I.I.D assumption* (Bengio, Lecun, and Hinton 2021). However, in complex applications, like automated fake news detection, it is unlikely that our test distribution resembles the real-world distribution. When we compute a standard accuracy score of a ML tool, we are only guaranteed this accuracy if the tool is deployed in a setting where the same distribution of inputs is given to it. This guarantee is not straight-forward when classifying complex, socially influenced, ever-changing items, like the veracity of information. Hence, we provide examples of where this guarantee is broken during deployment. The above two issues of bias and generalisability are not orthogonal. For example, evaluating all data points with equal weight neglects the inherent differences across the data points

(Hutchinson et al. 2022, Rodriguez et al. 2021), which can create uncertain predictions and create bias.

Our goal in this work is to highlight ethical challenges with deploying fake news detection models. This paper is formatted as such: we first describe the general approach to automated content moderation. Next, we review some broad ethical considerations in the context of algorithmic decision making. Building on these two foundations we put forth ethical considerations that must be taken into account in automated fake news detection. We then provide examples of key issues that arise from employing automated content moderation on over 140,000 news articles spanning one month in 2021. We do not base our analysis solely on the ground truth that the models were trained on. Instead, we expand our testing to articles from both inside and outside the training distribution to simulate what a real-world set of inputs may look like. We conclude with a discussion of our insights and avenues for future research.

2. Background

2.1. Automated content moderation

In the literature, the term 'content moderation' has been used to refer to both multiple outcomes and multiple targets, which are important to distinguish. Outcomes refer to the intervention that is triggered by a content moderation tool's prediction, whereas targets are the type of content being moderated by the tool. While many of the concepts examined in this paper can be applied broadly across both the outcomes and targets of moderation, our focus is on content moderation for targeting news articles (often called *fake news detection*).

First, content moderation can refer to both hard and soft outcomes. Hard content moderation refers to removing content. It can be done after the content has been published through the removal of a single piece of content or by banning the source of the content – known as Ex Post moderation (Jackson 2019; Katsaros, Yang, and Fratamico 2022). It can also be done before the content producer publishes information through post approvals or filters – known as Ex Ante moderation (Katsaros, Yang, and Fratamico 2022; Ribeiro, Cheng, and West 2022). Soft content moderation refers to interventions that do not remove content, but instead limit the visibility of content. Soft moderation can be done through attaching warning labels to content (Zannettou 2021), quarantining communities (Chandrasekharan et al. 2017; 2022), or demonetising content producers (Trujillo et al. 2020). Due to concerns of censorship, most researchers have moved towards soft methods (Zannettou 2021), though, recent research has suggested that using multiple types of moderation may be more beneficial than using just one (Bak-Coleman et al. 2022).

Second, the target of moderation can vary, but generally falls into a few categories: misinformation/disinformation, hate speech, terrorist/extremist content, spam, or pornography. Most current academic and industry work focuses on moderation for misinformation and hate speech, as they are still difficult to solve (Gillespie 2020; Kumar and Shah 2018; Zannettou et al. 2019). Further, many of the proposed techniques for moderating misinformation and hate speech overlap.

Within the work on misinformation and disinformation, our focus in this paper, there are two levels of data that fake news detection can act on: (1) data that is external to social media platforms, such as news websites, or (2) data that is internal to social media platforms, such as user-generated posts (Bozarth, Saraf, and Budak 2020; Reis et al. 2019; Shu et al. 2017). Across both data types, classifiers range in features used, algorithms used, and more. For instance, many classifiers have used text features – such as language-style in an article or user-generated post (Baly et al. 2019; Cruz et al. 2019; Hassan et al. 2017; Horne et al. 2018; Horne and Adali 2017; Potthast et al. 2017), relational features – such as relationships among news outlets or relationships between social media profiles (Gruppi, Horne, and Adalı 2021; Patricia Aires, Nakamura, and Nakamura 2019; Ruchansky, Seo, and Liu 2017; Shu, Bernard, and Liu 2019a; Shu, Wang, and Liu 2019b), and propagation features – such as diffusion network structures or temporal aspects of propagation (Metaxas, Finn, and Mustafaraj 2015; Resnick et al. 2014; Shu et al. 2020). Some of these feature sets have been handcrafted, while others have been automatically captured by neural network models. Further, learning algorithms have ranged from classic supervised algorithms (Horne, Nevo, et al. 2019; Reis et al. 2019), where lots of labelled training data is needed, to reinforcement algorithms (Mosallanezhad et al. 2022), where learning is done through exploration, to few-shot algorithms (Yang et al. 2019), which learn using little data.

Ultimately, what is common amongst these proposed tools, with some exceptions, is that they are built to classify content as *bad* or *good,* vis-à-vis some ground truth. While this may not be a literal bad/good scale, in most settings, this means that to train the classifier we need to label a set of data as fitting into two categories. For example, in news article classifiers, we could label the individual claims in the articles as being *true* or *false* (Hassan et al. 2017), we could label each full article as having *true* or *false* content (Horne and Adali 2017), or we could label each news outlet as being *reliable* or

unreliable (Horne, Nørregaard, and Adali 2019). There are classifiers that operate on more than two classes (Wang 2017), sometimes making groups such as *unreliable*, *mixed*, and *reliable*. There are classifiers that attempt to use scales rather than groups, such as factuality ratings from 0 to 100. Yet, no matter the method used, the line of what is considered *good* and what is considered *bad* must be drawn, and a choice of when to intervene in information consumption must be made. Drawing this line, in and of itself, can significantly influence future predictions by the model and hence, be an ethical issue to consider.

2.2. Ethical lenses

Content moderation involves three main actors: the content consumers, the algorithm, and the developers. Each of those actors might create- or be exposed to- bias, and each must make a judgement – by interacting with algorithmic advice, by using a feature model to make a prediction, or by choosing the ground truth to train a model with. To understand the challenges of content moderation, thus requires us to understand the inherent problems encountered by each. Exceedingly, consultation with scholars and practitioners familiar with the context of an ethical dilemma contributes to well-rounded, multidisciplinary, inclusive, and ethically sound solutions to such issues (Molewijk et al. 2004). Utilising this lens to understand the impact and inherent problems encountered by automated fake news detection provides a unique and innovative approach for our analysis.

Beginning with content consumers, we consider the notions of bias and genralizability through a utilitarian lens.[2] Utilitarian theory falls under the realm of consequentialist theories, for which the only indicator of the ethical status of an action are the consequences of that action. For utilitarians, the only thing with intrinsic moral value is happiness. Thus, when we have an ethical dilemma the right or ethical action is the one that maximises overall happiness or pleasure and minimises overall pain or suffering (Bentham 1843). Interestingly, truth in news does not have any inherent moral value for a utilitarian. Telling the truth would only be ethical if it maximised happiness, i.e. it led to more overall happiness than lying (Bentham 1843; Mill 2002). Similarly, bias in fake news detection would only be considered unethical if it caused more pain than pleasure, not because it is inherently bad. Commonly, two news outlets will report very differently on the same event. To label one or the other as good or bad is exceedingly political. For a utilitarian, all that would matter is how each of the outlets affected the promotion of happiness and the minimisation of pain. One's preference towards a more conservative or liberal telling of the event tends to influence our determination of the veracity of the information.

Utilitarian theory is embedded in the epistemological tradition of empiricism. For utilitarians, we obtain knowledge of the world through our senses. Consequently, when incoming information changes so does our understanding of what is considered to be true, and our ability to generalise from prior knowledge. This approach to knowledge requires that we regard the establishment of ground truth as an ongoing and contingent process. The fake news detection model would need to evolve with changing information in the world. Best practices, from this empiricist position, would require the maintenance of a feedback loop between the model's performance and the changing nature of news content. Finally, because bias would be perceived differently by different content consumers, a blanket use of content moderation models might not be perceived as ethical under this lens.

Considering the algorithms, with their heavy reliance on ground truth, we adopt Immanuel Kant's version of deontological ethics to elaborate on ground truth challenges. A deontological approach focuses on rules and duties in the evaluation of ethical dilemmas, consequences play no part. Kantian theory utilises, primarily, two formulations of his categorical imperative as well as an assessment of motive to assess the morality of action. One formulation of Kant's categorical imperative requires us to, 'act only on maxims that can become universal laws' (Kant 1998[1785]). In practice this means that when we are considering whether something is right or good, we must think about what would happen if that action became a universal law. What would happen if everyone did it? Kant's famous example is a lying promise. If you needed money and wanted to borrow it from a friend but knew you could not pay it back, would it be acceptable for you to lie to get the money? i.e. promise to pay it back even though you knew you could not. To find out whether the action is universalisable you must ask yourself what would happen if everyone took up this practice of lying to borrow money. What Kant concludes is that this results in a contradiction. When you universalise this action, you see that if everyone made these lying promises then everyone would also know that these promises were false and, thus, promise-making would lose its very meaning. Kant's second formulation of the categorical imperative requires that we act, 'in such a way that you always treat humanity … never simply as a means, but always at the same time as an end' (Kant 1998 [1785]). When we deny someone information (lying) and/or limit their freedom we treat them as a mean (object), which does not respect

their inherent moral value as a rational being. An excellent example of this is the concept of transparency as it is used in discussions of machine learning. If machine learning processes are unknowable, they are not transparent. If certain categories are used in machine learning, e.g. race in recidivism models, and this is not divulged to the very individuals that model outcomes affect then this categorical imperative is violated.

Kant's ethics is directly tied to his epistemology. As a rationalist Kant believed we could not rely on empirical observation to obtain truths about the world. The only things we can know for certain, without appeal to our senses, are things like mathematical truths (Shabel 2017). Otherwise, our cognitive composition affects the manner in which we understand the world. We filter information about the world through what Kant labels as our categories of understanding. We do not have direct access to the truth about the world. Kant's rationalist epistemology is an excellent reminder that all our perceptions of the world are biased, that our cognitive frameworks affect the way we understand facts about the world, and that generalisability of judgment may be limited. As we establish ground truths, we must keep this in mind and be open to understanding various perspectives on knowledge and truth. Thus, the establishment of ground truths should be viewed as an ongoing process.

Finally, we introduce virtue ethics to consider the actions of the developers of content moderation models. While Utilitarian and Kantian ethics focus on the nature of an action as integral to contemplating ethical status, virtue ethics takes a completely different approach and focuses on the development of particular character traits, i.e. virtues, in people. If a person develops all of their virtues properly, they will act virtuously (ethically). Simply, good people will do good things. Aristotle is well-known for his contributions to this ethical approach as the foundation of most modern virtue approaches. This theory proposes that we should all work to achieve a state of happiness and flourishing in our lives, something Aristotle termed as *eudaimonia*. To do this, we must practice perfecting the virtues that exist in all of us. Each virtue exists on a spectrum from excess to deficiency. On the middle of the spectrum the virtue exists in its perfect form, this midpoint on the spectrum is termed the golden mean. For example, honesty is considered a virtue. In excess, honesty can be a bad thing, e.g. being overly honest, in its deficient form you might lie. Thus, you must practice being perfectly honest. Once you have perfected all virtues you are a virtuous person and will always act virtuously. We could look at conflicts that might be created by virtues to understand challenges they present. As an example we will looks at ambition and fairness in the context of the leaderboard culture that has been embraced by the ML community. The leaderboard culture encourages programmers to create models that are increasingly accurate. Thus, an ambitious programmer will strive for such accuracy, possibly with a detrimental effect on society. However, the trade-off for accuracy in ML is, often, fairness (Kearns and Roth 2020), which is also a virtue. While the virtue of fairness is extrapolated to people affected by the model it is the programmer's excessive personification of ambition that leads to this. In a world that Aristotle could not have envisioned we can understand how a person's ethical responsibility might be somewhat disconnected from their action.

Closely tied to virtue ethics is virtue epistemology which, for some, suggests that our intellectual virtues, e.g. intellectual courage, open-mindedness, intellectual perseverance, and intellectual humility, are essential for gaining knowledge and forming justified beliefs (Zagzebski 1997). Zagzebski proposes that individuals who possess and cultivate these virtues are more likely to arrive at true beliefs and make intellectually responsible decisions. These intellectual virtues are vital in both the acquisition and dissemination of knowledge. Possessing these virtues enables individuals to effectively navigate complex epistemic situations, evaluate evidence, and arrive at well-founded conclusions. This perspective shifts the focus from merely evaluating the structure of beliefs to evaluating the character and intellectual virtues of the person holding those beliefs. Hence, for proponents of virtue ethicists and epistemologists, the burden lies with the individuals designing and implementing systems. This is crucially important to our paper's motivation – the people who are designing and implementing systems must understand their limitations and potential harm.

Building on the above, we recognise that the definitions and values pertaining to bias and generalisability of automated content moderation are relative to the ethical and epistemological frameworks of the designers and users of such tools. What one consumer might perceive as biased; another will accept as true. What one model may flag as unreliable; another will agree with. And what one developer may perceive as 'best' another may challenge. Because of this, we must clearly understand these issues before we implement them and rely on the judgement they offer.

3. Ethical considerations when deploying automated content moderation

In this section we return to focus on the two key challenges of deploying automated content moderation, namely bias and generalisability. We provide additional

background from the literature specific to each of these issues to better explain them. In the following section we provide empirical evidence of these issues across three different models from the literature. We show that these issues threaten the viability of fair and effective automated content moderation tools. We follow up with a discussion of potential solutions and alternatives.

3.1. When machines disagree, which one is correct?

Algorithmic bias refers to situations when the output of an algorithm benefits or disadvantages certain groups more than others without a justified reason for such unequal impacts (Kordzadeh and Ghasemaghaei 2022). It is commonly the result of algorithms picking up on the human biases of their designers (Kirkpatrick 2016). In the broader context of decision-making, biased algorithms are shown to hinder fairness, which is the absence of prejudice against individuals (or groups) based on their characteristics (Mehrabi et al. 2021), hence creating an ethical dilemma. There are many examples of such biases in the literature, perhaps the most famous is the recidivism algorithm shown to be racially biased. In the context of content moderation, bias may result in systematically flagging specific content as unreliable or favouring one source over another.

Algorithmic bias may arise from the data itself or from the design of the algorithm (Mehrabi et al. 2021) for example, through pre-existing bias that can affect the design of the system, technical bias that may arise during implementation, and personalisation – which may introduce further bias (Chouldechova and Roth 2020; Martin 2019). It can be difficult to disentangle what component of a tool is to blame for biased outcomes, as data and algorithm work together to make predictions.

Algorithmic bias can be measured in different ways. At the individual level, we expect that similar individuals would receive similar outcomes from the algorithm. Such measurement is contingent on our ability to properly quantify similarity (Kordzadeh and Ghasemaghaei 2022). At the group level, the former measurement approach translates to specific groups not being discriminated against. In content moderation, this approach would translate to similar articles (or groups or articles) being judged similarly, for example on their accuracy or reliability. Measuring bias here requires knowledge of *an objective ground truth* against which to compare judgement. Such ground truth will guide the algorithm to distinguish between true and false content, making training possible. However, as discussed in the previous section, an objective ground truth might not exist, or might change over time.

Taking a different approach to measuring bias, Cowgill and Tucker (2017) view algorithmic bias as a causal problem: if the introduction of a new algorithm causes outcomes to be more biased, then the algorithm can be seen as biased. Subsequently, they propose that it is possible to measure bias by studying disagreement between different algorithms. Such an approach can, to an extent, mitigate the impact that the ground truth has on measuring bias. In this paper we subscribe to this latter approach and measure bias not only vis-à-vis the ground truth, but also as comparative between different model judgements.

3.2. How far should we generalise?

While choosing the labels to train a system is important, ML systems do not simply regurgitate a set of training labels, rather they classify future, unseen, and unlabelled data based on the model. Simply put, if we imagine each data point in two-dimensional space, a ML system predicts what class a new, never seen before item is by comparing that data point to the labelled data points used during training (i.e. *how close is the new data point to data points labeled as good*?). Of course, most ML models use many features, making a high dimensional data space that is difficult to picture, but the idea is still the same.

Saliently, in this space, we can only make accurate predictions when the new data points given to the system are independent and identically distributed, which is known as the *I.I.D assumption* (Bengio, Lecun, and Hinton 2021). This means that ML models can only make reasonable predictions on data and situations similar to the past. While this assumption is fundamental to ML and predictive analytics, it is often ignored in fake news detection research. We assume that the fake news models can make extreme generalisations about veracity, reliability, and factuality, when, by definition, they can only make local generalisations. An illustration of this idea from Chollet and Allaire (2018) is shown in Figure 1.

The IID assumption may not only be broken when making predictions far outside the training data space, but it can also be broken when that space changes over time. In ML research, this is called a *distribution shift* – when the data distribution changes from training a model to deploying a model. A simple illustration of this idea is shown in Figure 2.

In general, distribution shifts have been well-studied in the broader ML community. In this literature, it has been shown that the IID assumption does not hold in

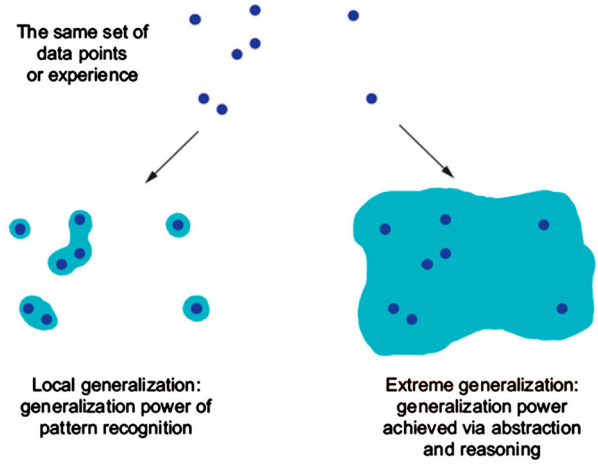

Figure 1. An illustration of local generalisation versus extreme generalisation from Chollet and Allaire (2018) *Deep Learning in R* book.

many practical deployment situations (Bengio et al. 2020), and that shifts in the data distribution can be difficult to detect and handle (Federici et al. 2021). Theoretically, these shifts can be due to the feature space changing (i.e. fake news no longer uses highly emotional language) or the concept of a data point changing (i.e. an unreliable news outlet is no longer unreliable) (Huyen 2022).[3] In practice, distributional shifts are often due to under or over-representing certain groups when training and testing a model, which may stem from the ground truth labelling scheme (Buolamwini and Gebru 2018). Yet, understanding and accounting for what sub-populations exist in the application's setting can be difficult. For example, in news classification, we could think of multiple overlapping and nested sub-groups of news that could impact the input distribution. For example, reliable news outlets may look different if they are focused on watchdog journalism or feature journalism rather than breaking news. News outlets may look different due to the country of origin or the sub-culture of their target audience. Opinion and analysis articles, which perhaps shouldn't be given as input to a veracity classifier, will certainly look different than articles that are breaking news. Further, these sub-groups may not only exist across outlets but within outlets. Hence, these variations would need to be represented and annotated for a fake news detection system to have any shot at correctly, and fairly, classifying veracity. Still, even if these variations can be fairly represented in an underlying data model, the evolving nature of news may still change the data distribution, making it difficult, if not impossible, to confidently rely on predictions.

To explore these two concerns further, we designed an empirical study that compares the prediction outputs of three fake news detection models proposed in the literature. Our goal with this empirical study is not to say which model is better, or to exhaustively compare all proposed models, but rather our goal is to provide concrete evidence of *what could go wrong* when deploying content moderation tools. We hope this exercise will provide a foundation for critically evaluating and questioning the tools we build. Further, we also hope that this exercise will encourage an *understand deeply* approach to content moderation research rather than a *just make something* approach.

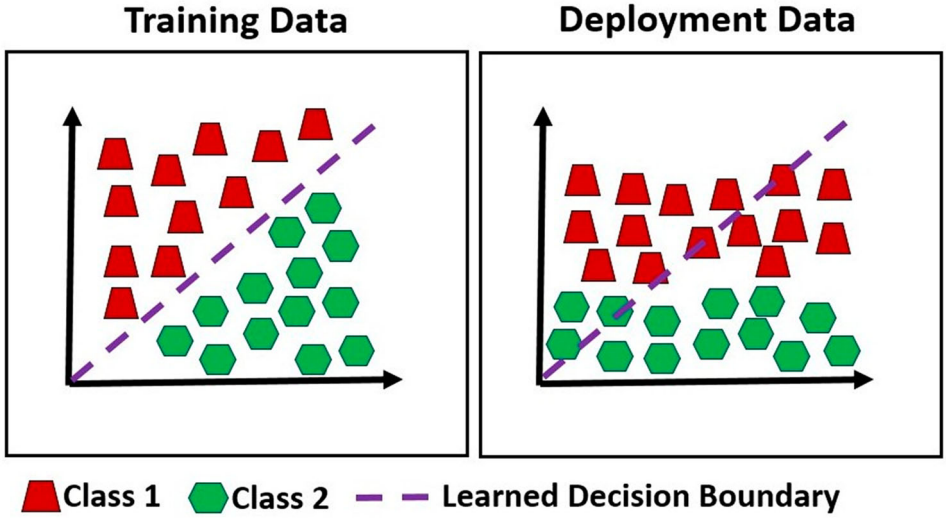

Figure 2. An illustration of distribution shift in ML applications. Our model may learn a decision boundary that is no longer valid when deployed due to data shifting.

4. Empirical evidence

To provide examples of the above two ethical concerns during the deployment of automated content moderation, we trained and analysed predictions from three different detection models. We chose three models that can be easily reproduced with publicly available code and are relatively explainable and transparent. While these three models only represent a fraction of the types of models and architectures proposed in the literature, the issues that emerge from deploying the models can point to more general problems. Furthermore, the ground truth labelling scheme used in training these three models is used across much of the literature.

In our case, given a news article, each model's task is to predict if the news article was published by a *reliable* or *unreliable* news outlet. For each feature model we trained a Random Forest classifier on the same set of over 381,000 news articles extracted from the NELA-GT-2021 dataset (Gruppi, Horne, and Adalı 2022a). These news articles were published by 270 outlets over 6 months (January 2021 to June 2021). Each model is trained on the same set of outlet-level ground truth labels from Media Bias/Fact Check (MBFC). These three models are further described in Table 1.

The NELA-GT datasets are yearly-released news article datasets from a wide range of high and low credibility news outlets, including legacy media (i.e. ABC News, Washington Post), hyper-partisan media (i.e. Newsmax, One America News Network), and conspiracy-driven media (i.e. Infowars, The Gateway Pundit). The datasets are often used in fake news detection modelling as they cover nearly every article published by each outlet during the year and the dataset comes with third-party credibility labels from MBFC. MBFC is a website that rates news outlets' factuality using a strict methodology to categorise outlets on a seven-point scale from v*ery high* factuality to v*ery low* factuality. In the NELA-GT-2021 data set, this seven-point scale is collapsed into three groups: *reliable*, *mixed*, and *unreliable* (Gruppi, Horne, and Adalı 2022a). This categorisation is done as parts of the MFBC seven-point scale are sparse. For model training and validation, we only use outlets in the *reliable* and *unreliable* categories. In general, labels from MBFC closely align with labels from other credibility rating systems, such as NewsGuard (Nørregaard and Horne 2019). We choose this method as both the setup and labelling scheme are commonly used in the literature (see examples in Baly et al. 2019; Ghanem et al. 2021; Gruppi, Horne, and Adalı

Table 1. Descriptions of each model used in our analysis, where accuracy is the average percentage of correct predictions on 20-fold cross-validation. As is done in literature, these folds are done at the source-level to match the ground truth: when an outlet is selected for testing, all articles from that outlet are left out of training. Note, the July 2021 test data set is not used when computing the accuracy of each model, only the cross-validation folds from the January 2021 to June 2021 data set.

Model	Model description	Labelling type	Accuracy on validation set
CSN	CSN is a supervised detection model based on a news outlets placement in a content sharing network, leveraging the idea that similar news outlets will copy content from each other more than dissimilar outlets (Gruppi, Horne, and Adalı 2021). Specifically, given a graph G = (V, E), where V is the set of news outlets and E are directed weighted edges representing the proportion of articles shared, each node is represented by its Node2Vec embedded vector (Grover and Leskovec 2016). This model has also been used in broader news embedding applications (Gruppi et al. 2022b).	Strong – Given a news outlet, the model predicts if that outlet is reliable or unreliable based on its placement in a content sharing network.	90%
NELA	NELA is a supervised, text-based model used across multiple studies (Baly et al. 2020; Barrón-Cedeno et al. 2019; Baly et al. 2018; Bozarth and Budak 2020; Horne et al. 2018; Horne, Nørregaard, and Adali 2019). NELA uses a hand-crafted set of 204 features from a news article's text and headline. These features capture writing style, writing complexity, bias language, moral-emotional language, and event-based language. NELA features have been used with a variety of supervised algorithms, but Random Forest is the most commonly used.	Weak – Given a news article, the model predicts if the article is produced by a reliable or unreliable outlet based on the individual article's features.	76%
BERT	BERT (Bidirectional Encoder Representations from Transformers) is a masked-language model built by Google for general natural language tasks (Devlin et al. 2018). BERT is widely considered a state-of-the-art baseline for natural language tasks and has been used in various ways (in whole or in part) in proposed fake news detection models (Heidari et al. 2021; Jwa et al. 2019; Kaliyar, Goswami, and Narang 2021; Kula, Choraś, and Kozik 2021; Lee, Liu, and Fung 2019; Singhal et al. 2019; Szczepański et al. 2021; Zhang, Harman, et al. 2020). We train a supervised detection model in which articles are represented by their average BERT sentence embedding.	Weak – Given a news article, the model predicts if the article is produced by a reliable or unreliable outlet based on the individual article's BERT representation.	73%

2021; Horne, Nørregaard, and Adali 2019; Patricia Aires, Nakamura, and Nakamura 2019).

We then tested each model with over 140,000 news articles from 550 outlets that were published in July 2021 (one month after the data that the models were trained on). These articles could come from any outlet that published a news article in July 2021, no matter if the outlet was labelled in the training set or not. This testing setup simulates what input to these classifiers may look like in a real-world deployment context (often referred to as *in the wild*). With these predictions, we created a data set that assigned a prediction (reliable/unreliable) from each model to each article, as well as the strength of prediction (probability of 'reliable'). Our final dataset contains 34,074 articles from 59 outlets labelled as *reliable* during training, 29,584 articles from 94 outlets labelled as *unreliable* during training, and 79,094 articles from 95 outlets that were not used during training. All news articles for testing were also extracted from the NELA-GT-2021 dataset (Gruppi, Horne, and Adalı 2022a).

As one part of our analysis, we also map news outlets in this test dataset to two other independent sets of labels that capture the political leanings of news outlets. The first set of political leaning labels is from Adfontes media. Adfontes media rates an outlet's political leaning on a seven-point scale from extreme left to extreme right. The second set of political leaning labels is from Allsides. Allsides rates an outlet's political leaning on a five-point scale from left to right. Both companies use different methods to develop these ratings. Anfontes has in-house analysts determine outlet ratings based on a predefine rating system, while Allsides uses a mixture of expert panels and surveys.

4.1. Algorithmic bias

We compared the (dis)agreement among the three models to identify potential bias. Specifically, we looked at articles that the strong-labelling and weak-labelling models disagreed on. Of 61,900 articles in which the models disagreed with each other, there were 25,466 (41%) that CSN predicted as coming from a reliable source, and 36,434 (59%) that CSN predicted as coming from an unreliable source. Within those CSN-reliable predictions, BERT and NELA both deemed 3,622 (14%) as unreliable. Within the CSN-unreliable predictions, BERT and NELA both deemed 17,389 (48%) as reliable. Thus, across these three models, there is very high disagreement. But in the absence of objective ground truth, which model is correct? Is it the model that is most accurate on the validation set?

Studying Figure 3, we can see the distribution of predictions made by each model (the three sets of charts) across outlet political leaning categories from two different media companies: Adfontes (top row) and Allsides (bottom row). In all charts, the horizontal axis reflects political leaning (left to right), and the vertical axis is the number of articles that fall under each category. The two colours represent the models' predictions of outlet reliability. From the distributions shown in Figure 3, the CSN model (which was the most accurate model on the validation set) predicts significantly more right-leaning outlets as unreliable than left-leaning outlets (greater portion of yellow compared to blue on the right-hand side of the charts). This bias is present, but much less so, in the weak-labelled, text-based models (NELA and BERT). For example, while, in our data, CSN predicts Fox News as an unreliable source, both BERT and NELA agree that numerous specific articles from Fox News are likely to have come from a sufficiently reliable source. Similarly, while CSN predicts The Huffington Post as a reliable source, both BERT and NELA deemed some articles coming from this source as unreliable. Neither Fox News nor The Huffington Post were outlets used during training.

This significant political bias from the CSN model is in part due to the model itself. The CSN model uses strong-labelling, meaning that each outlet is only represented by a single, constant feature vector. This setup means that the CSN will predict all articles from a single outlet the same way. Technically speaking, this setup is the most accurate at the task: predicting if an article is from a *reliable* or *unreliable* outlet. But this leaves very little room for nuance across articles. The feature space in the model is also responsible for this bias. As was shown in the paper where the model was originally proposed (Gruppi, Horne, and Adalı 2021) and as we will show further in Figure 5, right-leaning outlets are clustered significantly closer to conspiracy-peddling outlets in the CSN feature space than left-leaning outlets. Thus, in some way, this analysis simply points to a potential flaw in an individual model.

However, this bias is not all on the model but also the ground truth labels, which are used widely across different models. In Figure 4, we show the same political leaning categories from Adfontes and Allsides mapped to the ground truth labels that each model was trained on (reliability labels derived form MFBC). The x-axis is again the political leaning from left to right, while the y-axis is the number of outlets labelled as reliable (dark blue) or unreliable (dark yellow) in the training data. Across both independent political-leaning rating systems, we see that the outlet reliability training labels themselves are heavily skewed against right-leaning outlets, with 11 outlets labelled as right-leaning by Adfontes are labelled as unreliable outlets by MFBC, and 11

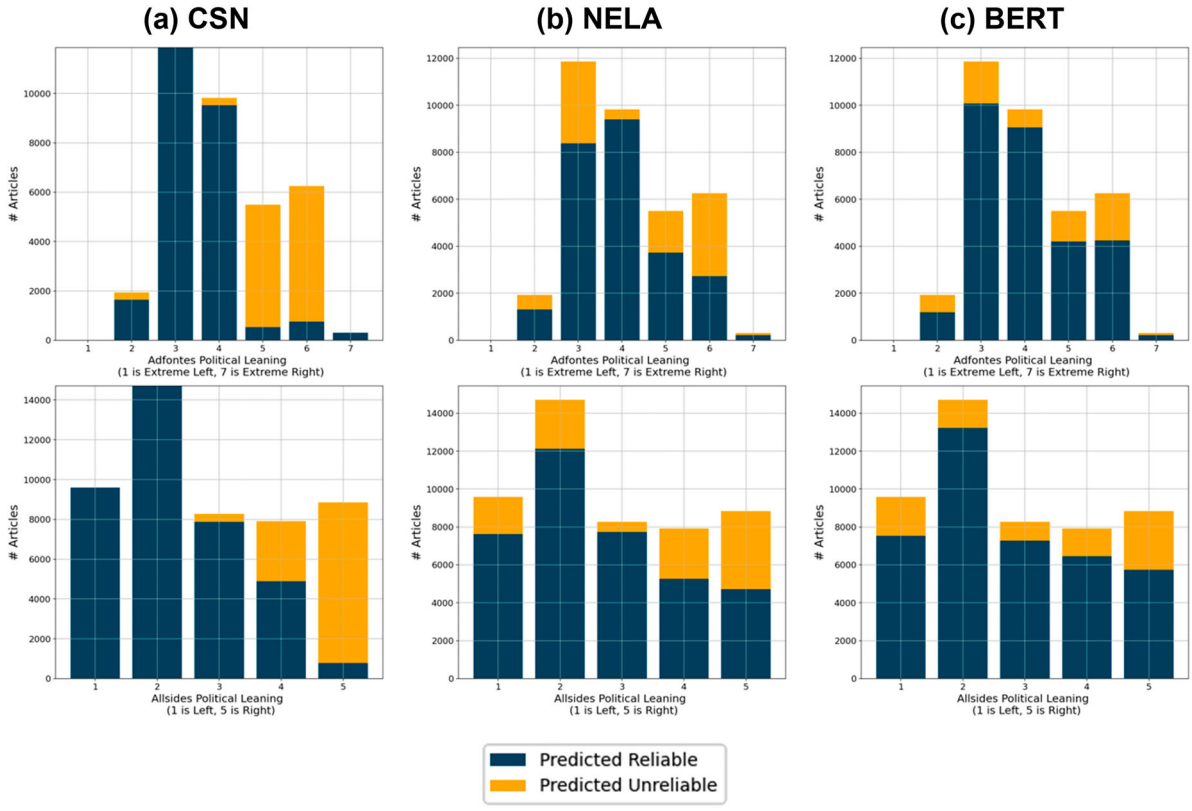

Figure 3. Distribution of predictions made by (a) CSN, (b) NELA, and (c) BERT across outlet political leaning categories from Adfontes Media (row 1) [https://adfontesmedia.com/static-mbc/] and Allsides (row 2) [https://www.allsides.com/media-bias]. Note, not all outlets are represented across the political leaning categories.

outlets labelled as right-leaning by Allsides that are labelled as unreliable outlets by MFBC. While only 3 outlets labelled as left-leaning by Adfontes were labelled as unreliable by MFBC and 1 outlet labelled as left-leaning by Allsides was labelled as unreliable by MFBC. While this skew may be in part due to the current media landscape, where it has been shown that false news is more often engaged with by conservative information consumers (Grinberg et al. 2019), that landscape may change over time and is likely not as heavily skewed

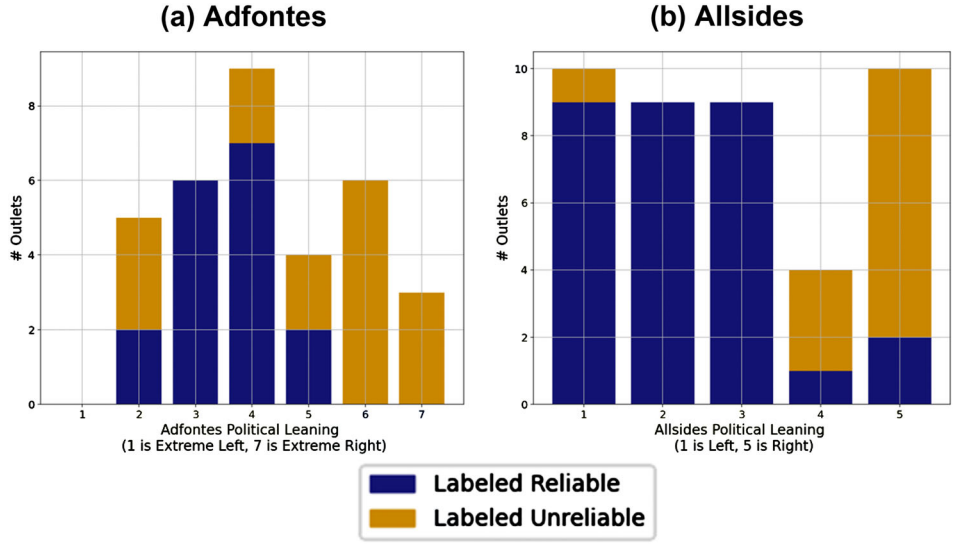

Figure 4. Distribution of outlet-level ground truth labels across political leaning categories from (a) Adfontes Media and (b) Allsides. Note, not all labelled outlets are represented across the political leaning categories.

as these labels indicated. The labels potentially changing over time is also a good example of a potential distribution shift when moving from training to deployment, as discussed in Section 3.2.

What we show in Figure 3 and Figure 4 is an example of aggregation bias (Mehrabi et al. 2021) in both the model space and the ground truth labelling scheme. Aggregation bias occurs when conclusions are drawn from populations and applied to individuals. While population-level analysis is a fast heuristic for predictions, for policy making, and to make sense of large amounts of data, we routinely err when we apply it, through induction, to individuals and ignore the uniqueness of each item's existence. Drawing conclusions based on population-level data may lead to false conclusions.

Automated fake news detection is prone to aggregation bias because, in many proposed fake news classifiers, ground truth labels stem from the reliability of a news outlet, rather than the truthfulness of an individual article. While this choice is made to operationalise the task (i.e. fact-checking is slow and selective, ML tools need large, timely datasets to make predictions), it may lead to false and/or unfair conclusions. On the other hand, training models on small, selective datasets of fact-checked news articles will likely lead to tools that struggle to generalise even more so than models that use outlet-level labels (we discuss such generalisations in greater depth in the following section). During ground truth creation, bias can occur at both the outlet-level and article-level. For example, since fact-checking is slow and selective, only highly engaged with topics and claims will be covered. Similarly, when labelling news outlets, only popular outlets may be labelled, and those determinations may be based on the veracity of articles on particular topics from those popular outlets. Both issues may create bias in the ground truth labelling scheme, whether drawing from population-level data or not.

We can also think about drawing conclusions based on population-level data in the feature space of fake news classifiers. Many text-based models utilise emotion in news articles to make predictions. Yet simply because the average fake news article in the past used highly emotional language does not necessarily mean that an individual news article in the future that uses emotional language is false. This reliance on emotional features may create a bias against investigative or watchdog journalism, which may report on situations with high degrees of emotion. The hope is that systems will not make erroneous predictions based on one feature in the model, but individual predictions outside of the traditional ML evaluation framework have rarely been explored.

4.2. Testing the IID assumption

Even if our ground truth labels for training a model are unbiased, they may not fully capture the deployment context. That is, by mapping a diverse set of data into two categories and leaving some data out of modelling, we may be ignoring the heterogeneity of the data, and our model may not reflect reality. In Figure 5, we show reduced feature spaces across each model for three example outlets: ABC News (labelled *reliable* during training), Breitbart (labelled *unreliable* during training), and Manchester Evening News (a U.K. news outlet that was never labelled during training). Each feature space was reduced using PCA. What is clear from these plots is that the labelled outlets (blue and red) look very different than the unlabelled outlet (green),

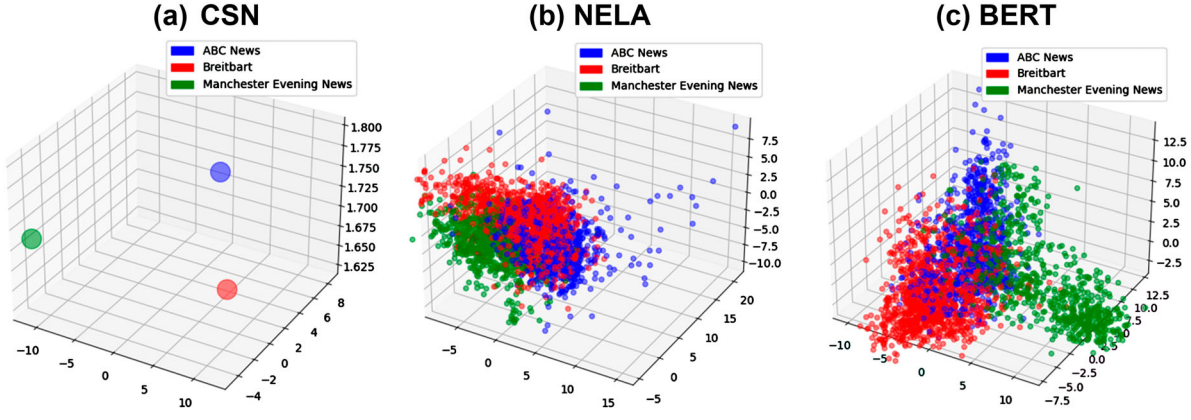

Figure 5. Three-dimension PCA Plots of the feature space for 3 outlets in each model. ABC News is labelled as reliable in training, Breitbart is labelled as unreliable, and Manchester Evening News is unlabelled. Note that the unlabelled outlet data points look different than both labelled classes – particularly in the (a) CSN feature space and the (c) BERT feature space. The CSN model only has 1 data point for each outlet, while BERT and NELA have data points for each article.

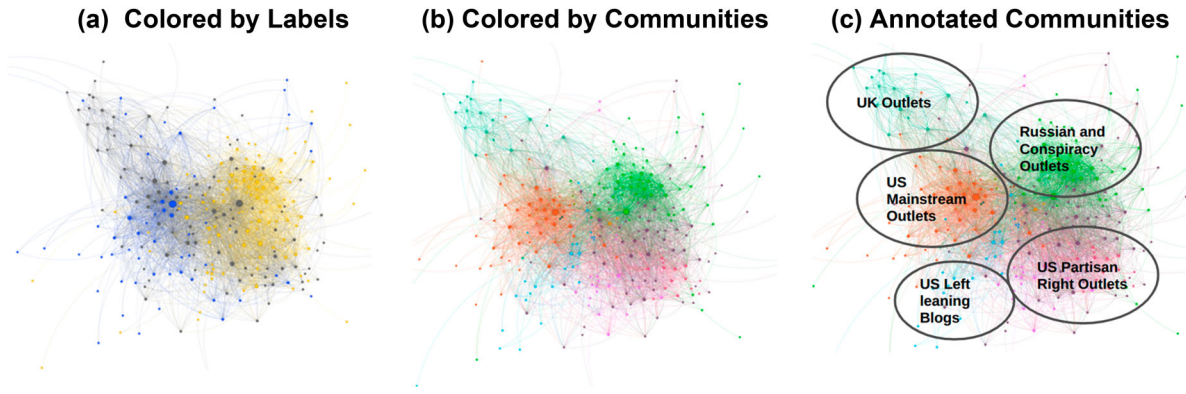

Figure 6. To better illustrate the distribution differences between outlets in the CSN, we show the content sharing network that underlies the model. In (a), we show the network with nodes colour by the outlet labels used during training, where blue is *reliable*, yellow is *unreliable*, and grey is unlabelled. In (b), we show the same network with nodes coloured by community membership, as determined by modularity. In (c), we show the same network with annotated with general community descriptions.

particularly in the CSN and BERT feature space. If an analyst were to look at this feature space without knowing the labels, they might assume there are three categories of data in the model, when in fact there are only two. That high dissimilarity creates predictions that are outside of the range of the training data, which may not be valid like predictions made within the range of the training data.

To further illustrate how far we are attempting to generalise when predicting the reliability of certain outlets (like U.K. news outlets) in the CSN model, we show the underlying content sharing network used in training the CSN in Figure 6. In Figure 6a, we can see that the reliable and unreliable labelled outlets form clear community structures, where blue is reliable and yellow is unreliable. This neat clustering is the reason the CSN model performs well on validation sets. However, if we are given a news outlet that does not appear in or close to these well-defined communities, it is uncertain what classification will be given. In the case of our test data, the U.K. news outlets form a community that is far from both the reliable and unreliable labels (Figure 6b,c).

When we examine specific disagreements in the data, uncertainty and potential biases when predicting outside the training distribution emerge. For example, when predictions are made on Sputnik News, a Russian state-owned news outlet that is well-known for its part in disinformation campaigns and is not used when training the models, we see mixed results. CSN predicts that Sputnik is an unreliable outlet, but NELA predicts that 86% of the articles from Sputnik are from a reliable source, and BERT predicts 81% of the articles from Sputnik are from a reliable source.

Problems when generalising can also be seen in predictions made on outlets that are used when training the models. For example, The Daily Mail is generally perceived to be an unreliable source and is labelled in training as an *unreliable* source. CSN indeed gives it a very low probability of being reliable (0.08). However, out of 204 articles from the Daily Mail in our test data set, BERT and NELA both predicted 142 of them (~70%) to be from a reliable source, and for an additional 54, either BERT or NELA predicted that the article is from a reliable source. Another example is when predictions are made on articles from The Root, an African American-oriented news outlet that is labelled as *reliable* during training. NELA predicted that 73% of The Root's articles are from an unreliable source, while BERT predicted 15% of The Root's articles are from an unreliable source.

Cultural differences in writing style can diminish the ability of text-based models, like BERT and NELA, to accurately score articles outside of the US mainstream – as most of the labelled reliable outlets are from the US mainstream. This again demonstrates the abovementioned generalisability issue and is supported by prior literature (Gruppi et al. 2018; Horne, Nørregaard, and Adali 2019). This issue also suggests that prediction uncertainties can emerge not only when predicting far outside the training data set, but also when there is high heterogeneity in the underlying feature space of the training data. Despite both the Daily Mail and The Root being labelled in training, their predictions in the wild were, for the most part, the opposite of their assigned ground truth labels. This illustrates that distributional shifts and over-generalisations in the fake news detection space are multi-faceted and difficult to define clearly.

4.3. The news cycle breaks model assumptions

There are other ways that the I.I.D assumption can break through distributional shifts or out-of-

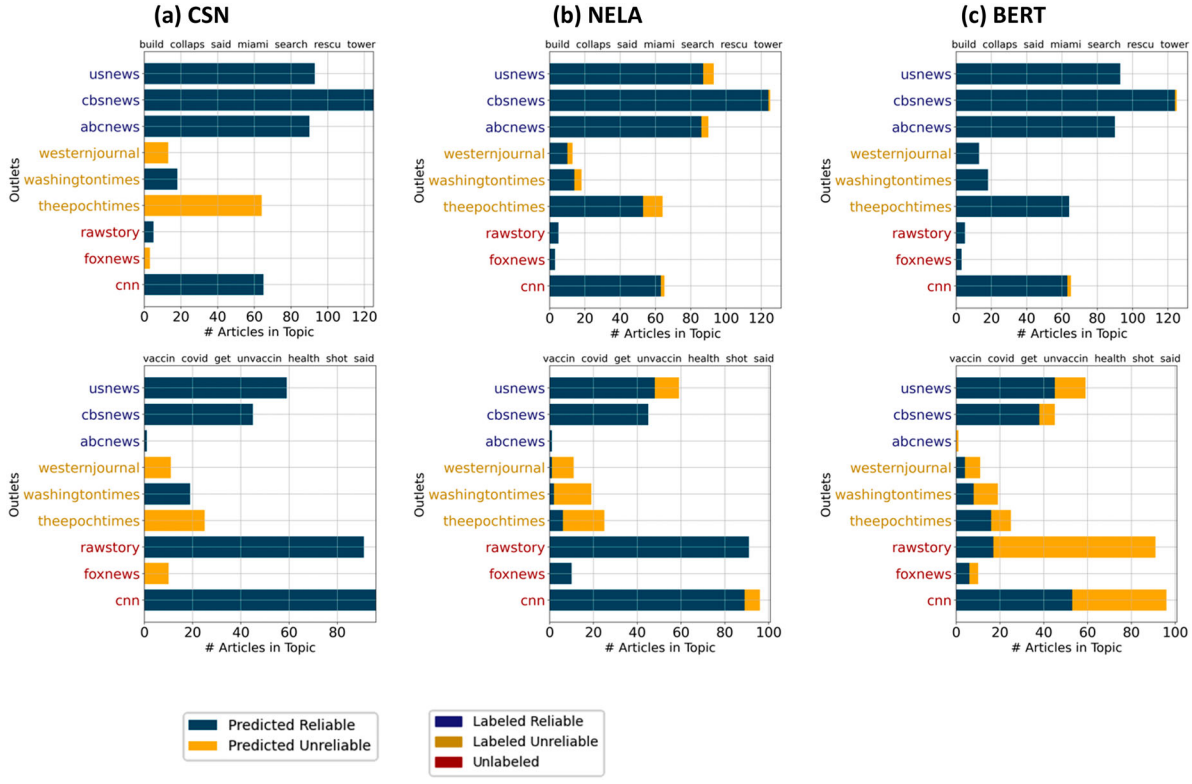

Figure 7. Distribution of predictions made by (a) CSN, (b) NELA, and (c) BERT on articles from 9 selected outlets in two topics: Surfside condominium collapse (row 1) and COVID-19 vaccinations (row 2). Each plot is titled with the most probable topic words from the model. Outlet names on the y-axis are coloured by their label in training the models, where blue is *reliable*, yellow is *unreliable*, and red is unlabelled/not used in training. Bars in the bar chart are coloured by the number of articles from an outlet predicted as *reliable* (blue) and *unreliable* (yellow).

distribution generalisation in the wild. While distributional shifts are difficult to define clearly, the news cycle itself can reveal some of these shifts. Media is constantly changing in reaction to events and the public, meaning that news topics alone may change how a feature space looks. Hence, even if we somehow can properly annotate and account for the various sub-groups discussed above, the feature space may still change over time simply due to topics in the news cycle changing. This property of news may make predictions in the wild *unpredictable* and make fake news detection prone to errors that are not captured in traditional evaluation frameworks (Bozarth, Saraf, and Budak 2020; Horne, Nevo, et al. 2019). From an ethical standpoint, this notion is similar to Kant's lying promise example discussed previously in this paper. If the model purports to predict what is unpredictable, then we end up with a contradiction. To understand just how predictable news reliability *actually* is in the wild, researchers and practitioners need to move beyond 'learner-centric' evaluation frameworks (Hutchinson et al. 2022).

An example of this behaviour is shown in Figure 7. In Figure 7, we show predictions made by each model across two different topics: (1) the Surfside condominium collapse in Miami, Florida and (2) COVID-19 vaccinations. Articles were grouped into topics via a Structured Topic Model (STM).[4] We show these predictions over articles from nine selected outlets: three outlets labelled as *reliable* during training, three outlets labelled as *unreliable* during training, and three outlets never labelled during training. What is apparent is that, given the model's task of predicting if an article is from a reliable outlet or an unreliable outlet, the text models do not predict consistently across topics, thus creating additional bias. For instance, BERT predicted that 99.2% of the articles produced by CBS News (an outlet labelled as *reliable* during training) were from a reliable outlet when the topic was the Surfside condominium collapse. Yet, the same model predicted that only 84.4% of articles produced by CBS News were from a reliable outlet when the topic was COVID-19 vaccinations. Likewise, BERT predicted that 98.4% of articles by CNN (an outlet not labelled during training) were from a reliable outlet when the topic was the Surfside condominium collapse and only 55.2% of articles were from a reliable outlet when the topic was COVID-19 vaccinations. Similar shifts in

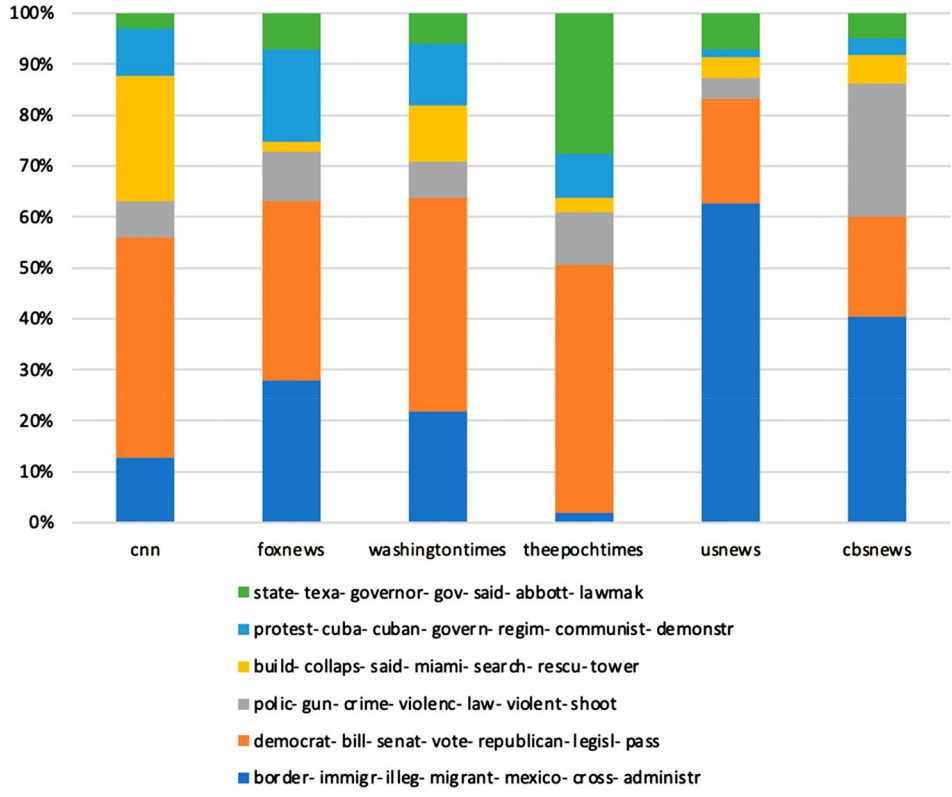

Figure 8. Illustration of the distribution of a sample of topics across six sources. Two outlets from each of our labelled and unlabelled groups: US News and CBS News were labelled as *reliable*, Washington Times and The Epoch Times were labelled as *unreliable*, and Fox News and CNN were unlabelled sources. Note that each outlet, even within the same label group, produces a different number of articles across topics. The relationship between outlet and topical coverage is significant across the data set.

the predictions from BERT happened across all 9 outlets, where articles on the condo collapse were more often predicted as reliable (473 articles out of 476 total articles in the topic) and articles on COVID-19 vaccination were more often predicted as unreliable (170 articles out of 357 total articles in the topic). This pattern also holds true for NELA but much less extreme than BERT.

Note, due to the CSN model not utilising individual article text, the feature distributions cannot shift due to topic changes, as the outlet placement in the network remains constant. This consistency is one of the benefits of the CSN model, but as discussed earlier, it may perpetuate political biases from the training labels more than the weakly-labelled text models (NELA and BERT).

The imbalance in topical coverage may also be impacting fake news detection models during training. There may be interactions between source and topic in a way that resembles confirmation bias (Nickerson 1998). That is, sources may report differently on topics if they align with the source's orientation (Eilders 2000; Iyengar and Hahn 2009), which in and of itself could be considered information decontextualisation by a news outlet. This also means that our training data may represent some topics more than others, and those imbalances very likely change across training labels. To concretely show this issue, we considered the news coverage offered by each source, by topic in our test data set. In a Chi-Square test of independence between news outlets and topics, our data showed a clear and significant relationship between news outlet and topics. Such relationship implies a representation bias (Mehrabi et al. 2021), which arises from biased sampling from the population. This relationship also, again, points out the high heterogeneity of the data that is ignored in training. A partial example of what this looks like in our dataset is presented in Figure 8.

5. Discussion

In this work, we focused on two key examples of ethical issues that can emerge when using fake news detection in the wild. First, we demonstrated that classifiers may be biased against specific news outlets due to both model design and the choice of training labels. Accounting for and handling these biases during model evaluation is not straight-forward. Second, predictions in

the wild may be *unpredictable* due to the I.I.D assumption being broken. We provide examples of this core assumption being broken by making predictions across cultures and countries, by not accounting for latent subgroups in the data, and through topical changes in the news cycle. From these specific examples, we see three big picture issues in the fake news detection space: 1. Who chooses the ground truth matters, 2. Operationalising tasks for automation may perpetuate bias, and 3. Ignoring or simplifying the application context reduces research validity.

5.1. Who chooses the ground truth matters

As has been suggested by a prior study (Bozarth, Saraf, and Budak 2020a) and by the above case studies, ground truth labels can significantly vary model outcomes, making the choice of what is *good* and what is *bad* in model training critical. Despite this criticality, for the most part, model evaluations by academic researchers depend completely on labels made by a third-party, such as Media Bias/Fact Check, NewsGuard, Ad Fontes Media, Snopes, opensources.co, or GossipCop. While understandably researchers do not want their own subjectivity to impact the labelling process, how these organisations decide what to label and how to label it may lack transparency and have its own biases. In fact, some of these credibility labellers are for-profit companies who have proprietary systems. If a classifier is later deployed on a for-profit media platform, such as Facebook, Twitter, or TikTok, another layer of opaque decision-making on what is considered good and bad may be added.[5]

This notion of who decides on ground truth labels points to a broader ethical issue – outside of the issues of process transparency and conflicts of interest. Namely, by training a model on labelled data, the assumption is that the labels are 'objectively singular and knowable', contrasting the reality that information may be 'socially and culturally dependent' (Hutchinson et al. 2022). The labels are embedded in the model and the model may be used widely across information situations and contexts. As social and cultural norms, worldviews, and socio-economic contexts influence information processing and perceived realities (Cronk 1999; Durkheim 1915[1965]; Lewandowsky, Ecker, and Cook 2017; Newman, Nisbet, and Nisbet 2018), and those realities can change over time (Bentley, Hahn, and Shennan 2004; Varnum and Grossmann 2017), who chooses what is *good* and *bad* influences the model, as their own cultural background is embedded into that model. In this way we are 'imposing hegemonic classifications' (Prabhakaran, Qadri, and Hutchinson 2022). This idea is more generally articulated by Prabhakaran, Qadri, and Hutchinson (2022):

> Since language and symbols, and ontology and axiology, play a critical role in the development of AI systems – e.g. through 'labels' on data, and how 'knowledge', 'objectivity', 'reality/truth', and 'system objectives' are constructed – the cultural norms of the AI developers and researchers also pervasively infuse the AI systems.

So then, who should determine the ground truth when training a fake news detection system? The developers? For-profit media companies? American Journalists? Researchers need to think critically about how to incorporate a range of cultural perspectives into the training pipeline or if such a complex advancement is even possible (Prabhakaran, Qadri, and Hutchinson 2022). One must also consider the ethical lenses previously discussed and how they might affect the determination (and conceptualisation) of ground truth.

5.2. Operationalising tasks for automation may perpetuate bias

In multiple applications of ML, complex tasks have been simplified for automation. By 'oversimplifying' a complex task to be automated and reducing the systems deployment to only technical questions, it is likely that we will fail in practice (Dahlin 2021). As a concrete example from outside the content moderation space, the 'Optum' algorithm was designed and implemented by a healthcare company to identify and prioritise patients for extra care. A study found that the algorithm, which used data such as past healthcare utilisation and costs to predict which patients are most likely to benefit from additional care, exhibited significant racial bias (Obermeyer et al. 2019). Specifically, the algorithm assigned a higher level of risk to white patients compared to equally sick black patients, leading to black patients being under-prioritised for additional care. In an example of oversimplifying the complex task of prioritising patients, the algorithm used healthcare costs to predict healthcare needs, assuming that sicker patients will spend more on healthcare. However, less money is typically spent on black patients with similar health profiles as white patients (Obermeyer et al. 2019), ultimately resulting in bias.

The ground truth labels, seemingly objective categories, end up reflecting and perpetuating bias in the system. This is akin to the healthcare algorithm's training on data that reflected seemingly objective, accurate, and benign facts. The data set used for training was

comprised of a patient population in which the average yearly dollar amount spent on healthcare for black and white patients was equal. This seems fair and unbiased but it ignores the fact that, on average, when a black patient and a white patient spend the same dollar amount on healthcare, the black patient is sicker. 'Facts' are not always as straightforward as they appear. As our analysis demonstrated, even when we think we have justification to label something as a ground truth we must challenge ourselves to unpack these notions further until we are certain that we are not using compound 'truths' as proxies for simpler notions.

Operationalising the task of predicting information veracity using outlet reliability is somewhat analogous to using current healthcare expenditures as a predictor of future healthcare costs. In fake news detection, we create categories and labels that inadequately capture the complex phenomena of mis- and dis-information (Wardle and Derakhshan 2017), lumping many concepts into a single, simplified framework. This label is being used as a proxy for other, more accurate and informative indicators. Although using more accurate, granular, and informative labels may be difficult to annotate at scale for training a classifier, those more complex labelling schemes may better capture the reality of the deployment situation.

5.3. Ignoring or simplifying the application context reduces research validity

As has been more broadly argued by Hutchinson et al. (2022) and Bengio, Lecun, and Hinton (2021), current ML evaluation frameworks do not always translate to practice. When we focus our efforts on beating narrowly defined state-of-the-art benchmarks, the validity of our research weakens, and perhaps more importantly, transferring that research into practical technologies can be dangerous. The application context must be fully considered. We assume that these systems can be applied broadly and make extreme generalisations, when by the very rules of predictive analytics, they can only predict in limited contexts. Hence, claiming that a fake news classifier has 90% accuracy may mean nothing if that classifier was used in real-life (the application context). If classifiers are going to be used in practice, finding ways to limit the scope of their predictions is critical.

To limit the scope that these tools predict in, we must study when they are the most effective and safe. Traditional ML evaluation frameworks treat all errors as the same, even though different errors may impact consumers differently than others (Hutchinson et al. 2022). If we can more concretely understand the cost of errors on the end-users, we can better understand what mistakes can be tolerated when building automated moderation systems. A simple example of this idea comes from page 29 of Yaser, Magdon-Ismail, and Lin (2012):

> Consider two potential clients of [a] finger print system. One is a supermarket who will use it at the checkout counter to verify that you are a member of a discount program. The other is the CIA who will use it at the entrance to a secure facility to verify that you are authorized to enter that facility.
>
> For the supermarket, a false reject is costly because if a customer is wrongly rejected, she may be discouraged from patronizing the supermarket in the future. … For the CIA, a false accept is a disaster. An unauthorized person will gain access to a sensitive facility.'

The choice of error measure depends on how the system is going to be used. However, in the case of content moderation, we do not have a clear picture of what the costs of different types of errors are. For example, if a U.S. left-leaning information consumer sees a news article from a left-leaning outlet incorrectly labelled as false (a false reject), how does that consumer react? Do they distrust the tool in all future interactions (where it may be correct more than it is incorrect)? Does this create *backfire effects* down the road (Swire-Thompson, DeGutis, and Lazer 2020; 2022)? Or perhaps one mistake doesn't matter, as the user will learn to trust the tool over multiple correct interactions. The nuances and variations in these reactions matter, particularly if a tool is deployed to a large audience. To define our error measures and when our tools should be allowed to make predictions, we must carefully study the cost of making mistakes during deployment.

6. Conclusion and future work

The issues of establishing and operationalising ground truth in model training and evaluation, raised above, deserve significant attention. There are at least two ways in which we can look to ethics and philosophical theory to help us solve this problem: multidisciplinary inclusion in decision-making (Molewijk et al. 2004; Beecher 1966) and Kuhnian paradigms to ensure ongoing re-establishment of ground truths (Kuhn 2012).

It is necessary to have a diverse and multidisciplinary team to assess the ethical concerns that inevitably arise in AI model training. This approach has been adopted by many disciplines and contributes to a more rigorous and holistic evaluation of issues and the development of creative solutions (Herkert 2005; Kitto and Sylvester 2002). AI systems have far-reaching societal, economic, and ethical implications that cannot be adequately understood or resolved by a single field alone. By

drawing on the insights and methodologies of various disciplines, we can arrive at more robust, balanced, and effective solutions that consider the diverse dimensions of ethical dilemmas.

Developing and deploying machine learning and artificial intelligence products results in consequences that affect the population in domains such as economics, sociology, psychology, medicine, legality, etc. When working on the determination of ground truth there is certainly a need for experts in technology, computer programming, data analytics, etc. That is to say, we need experts who can understand and explain the inner workings of the AI but we also need professionals of other academic fields and members of the general public. Increasing the diversity of perspectives allows us to consider more possibilities in terms of positive and negative consequences. AI ethics is not a problem limited to one discipline but a web of interconnected issues. A multidisciplinary team can address the topic holistically, identifying the root causes, potential consequences, and unintended effects that might arise across various domains. This prevents solutions that might inadvertently create new problems in other areas. Such an approach might also have the benefit of creating public trust in AI. The transparency created by, for example, involving the public, will allow the public to understand and be involved in the deployment (Robinson 2020; Züger and Asghari 2023).

Review of deployed ML applications must be ongoing. However, this does not mean that every single anomaly requires a complete overhaul of the application. Thomas Kuhn's seminal publication *The Structure of Scientific Revolutions,* which provided an account of scientific progress, can provide insights into the ongoing establishments of ground truths. Kuhn's approach to scientific frameworks emphasises the idea of scientific revolutions, where new paradigms replace old ones through a process of crisis, anomaly recognition, and paradigm shifts (Kuhn 2012).

For our purposes, we can think about our initial determination of ground truth as a 'normal period'. We use current evidence to determine what we consider to be true. Initially, at least, this might go well, there may be instances where the current models fail to perform as expected or situations where the ground truth is uncertain or disputed. Eventually, as these anomalies accumulate, questioning of the existing paradigm begins and we start to seek new ways to establish ground truth. At this point the existing paradigm faces a crisis. This can trigger a paradigm shift, where researchers explore new approaches and frameworks for establishing ground truth. For example, if a dominant AI model fails to generalizse well to certain scenarios, it could prompt the exploration of alternative data labelling methods or more robust model evaluation techniques.

Proposed new methods for establishing ground truth now need to be rigorously tested, validated, and accepted by the AI community, which includes multidisciplinary representation. This process involves peer review, empirical evidence, and consensus-building. These new methods may lead to advancements in the field and might be integrated into the larger AI framework, potentially altering how AI models are trained, evaluated, and deployed. Kuhn's theory emphasised the cyclical nature of scientific progress, where paradigms shift and evolve over time. Similarly, the methods for establishing ground truth are likely to, and should, continue evolving as the field progresses, responding to new challenges and discoveries.

This paper only highlighted a few of the many potential problems in automated content moderation. Just as described by Char, Abràmoff, and Feudtner (2020) in their discussion of ML for healthcare applications, the uncertain impacts of emerging technologies present a barrier to building an ethical framework. While we can theorise and construct guidelines for deploying such tools, these frameworks remain incomplete without studying the impacts of emerging technologies. Hence, one of our hopes is that this work will call attention to studying the impacts of content moderation tools in controlled environments in order to better understand the potential negative impacts and better build ethical frameworks, moving beyond leaderboards and static benchmarks.

Furthermore, no matter how accurate a model is evaluated to be, mistakes will be made. When a weather forecast predicts there will be a 90% chance of rain, the 10% chance it does not rain has little effect on the forecast's consumers: *we carry around an umbrella and never have to use it*. However, when a fake news detection tool predicts that there is a 90% chance a news article is reliable, what happens if the low-odds classification comes true? We make strong assumptions that abstract concepts can be mapped cleanly onto well-defined categories. And we assume that the humans interacting with model outcomes understand this mapping. Thus, understanding the potential harm to information consumers caused by an automated tool making a mistake becomes important; perhaps more so than evaluating an individual model's performance. As argued in the discussion section above, understanding the benefits and harms of deploying automated content moderation requires us to study information consumers' interactions with and reactions to these interventions, both when those tools work well and when they err.

Finally, we must be open to the idea that automating content moderation may have limited-use, and instead find alternatives. Again, to properly answer this question we must study it. For example, is it more effective, fair, and safe to broaden the reach of media literacy training instead of widely deploying automation? Is it more effective, fair, and safe to deploy 'misinformation inoculations' from Psychology (Van der Linden et al. 2017) instead of widely deploying automation? Is a combination of media literacy and limited-use automation the most effective and safe strategy? Empirically comparing these proposed solutions in their application context will require work across disciplines, rather than each solution being siloed in its own field. Our hope is that this work encourages interdisciplinary studies of proposed misinformation solutions and less reliance on limited technical evaluations.

Once individual methods for disinformation mitigation (automated and not) are critically evaluated and compared, we can begin to be creative with designs and control sequences. While these designs should be dependent on the results of the other studies, we can theorise about the types of designs that are possible. Automation may still play a role in content moderation, however a limited role. And this limited role can be supplemented by a safe, non-automated alternative. As argued in this paper, we often assume that automated systems can be applied broadly and make extreme generalisations, when by the very rules of predictive analytics, they can only predict in limited contexts. If classifiers are going to be used in practice, finding ways to limit the scope of their predictions is critical. As our analysis suggests, models generalizse poorly when trained on and deployed across all types of news. However, we may be able to build reasonable classifiers for specific types of news and construct automated pipelines to limit when these classifiers are allowed to make predictions. For instance, if we trained a classifier to predict if U.S. political news is reliable, we could utilise simpler classifiers for topic and country of origin to filter the input given to the veracity model, thereby limiting what types of news the tool can make predictions about. We can limit the model even further by only allowing warning labels to be attached to posts only when the tool is highly confident in the result. This limited-use of automation can be paired with a non-automated solution like news literacy reminder labels that appear on all the posts that the classifier does not make predictions on. By combining weak, limited solutions, we may be able to create strong, robust, fair, and safe solutions. Additionally, automated techniques may be more feasible if built to target specific types of disinformation, rather than attempting to ascribe a one-size-fits-all automated solution.

Notes

1. Computed using https://zenodo.org/record/1218409\#.YrHv_NLMKrw
2. Note that we don't mean to imply that a specific lens dominates others, rather we discuss each specific lens as an example to demonstrate potential dangers of blindly applying content moderation models and tools.
3. See Chip Huyen's 2022 blog post for an easy-to-read explanation of distribution shifts: https://huyenchip.com/2022/02/07/data-distribution-shifts-and-monitoring.html#concept-drift
4. A STM is a generative model of word counts that allows for the use of document-level metadata, commonly used to extract themes from text data (Roberts et al. 2014).
5. The hope is that with self-regulatory standards, such as those from the European Commission in 2022 (https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation), would prevent harmful opaque decision making from companies doing content moderation.

Disclosure statement

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

ORCID

Benjamin D. Horne http://orcid.org/0000-0002-5779-3019

References

Allcott, H., and M. Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31 (2): 211–236.

Asatiani, A., P. Malo, P. R. Nagbøl, E. Penttinen, T. Rinta-Kahila, and A. Salovaara. 2020. "Challenges of Explaining the Behavior of Black-Box AI Systems." *MIS Quarterly Executive* 19 (4): 259–278.

Bak-Coleman, J. B., I. Kennedy, M. Wack, A. Beers, J. S. Schafer, E. S. Spiro, and J. D. West. 2022. "Combining Interventions to Reduce the Spread of Viral Misinformation." *Nature Human Behaviour* 6 (10): 1372–1380.

Baly, R., G. Karadzhov, D. Alexandrov, J. Glass, and P. Nakov. 2018. Predicting Factuality of Reporting and Bias of News Media Sources. *arXiv preprint arXiv:1810.01765*.

Baly, R., G. Karadzhov, A. Saleh, J. Glass, and P. Nakov. 2019. Multi-Task Ordinal Regression for Jointly Predicting the Trustworthiness and the Leading Political Ideology of News Media. *arXiv preprint arXiv:1904.00542*.

Baly, R., G. D. S. Martino, J. Glass, and P. Nakov. 2020. We Can Detect Your Bias: Predicting the Political Ideology of News Articles. *arXiv preprint arXiv:2010.05338*.

Barrón-Cedeno, A., I. Jaradat, G. Da San Martino, and P. Nakov. 2019. "Proppy: Organizing the News Based on Their Propagandistic Content." *Information Processing & Management* 56 (5): 1849–1864.

Beecher, H. K. 1966. "Ethics and Clinical Research." *New England Journal of Medicine* 274 (24): 1354–1360.

Beltramin, D., E. Lamas, and C. Bousquet. 2022. "Ethical Issues in the Utilization of Black Boxes for Artificial Intelligence in Medicine." In *Advances in Informatics, Management and Technology in Healthcare*, edited by John Mantas, P. Gallos, and E. Zoulias, 249–252. IOS Press.

Bengio, Y., Y. Lecun, and G. Hinton. 2021. "Deep Learning for AI." *Communications of the ACM* 64 (7): 58–65.

Bentham, J. 1843. *The Works of Jeremy Bentham*. Vol. 7. Edinburgh: W. Tait.

Bentley, R. A., M. W. Hahn, and S. J. Shennan. 2004. "Random Drift and Culture Change." *Proceedings of the Royal Society of London Series B: Biological Sciences* 271 (1547): 1443–1450.

Birhane, A., P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao. 2022. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 173–184.

Bowman, S. R., and G. E. Dahl. 2021. What will it take to fix benchmarking in natural language understanding?. *arXiv preprint arXiv:2104.02145*.

Bozarth, L., and C. Budak. 2020. "Toward a Better Performance Evaluation Framework for Fake News Classification." In *Proceedings of the International AAAI Conference on Web and Social Media*, 14: 60–71.

Bozarth, L., A. Saraf, and C. Budak. 2020, May. "Higher Ground? How Groundtruth Labeling Impacts our Understanding of Fake News About the 2016 US Presidential Nominees." *Proceedings of the International AAAI Conference on Web and Social Media* 14: 48–59.

Buolamwini, J., and T. Gebru. 2018. *Conference on Fairness, Accountability and Transparency*, 77–91. PMLR.

Carter, B., S. Jain, J. W. Mueller, and D. Gifford. 2021. "Overinterpretation Reveals Image Classification Model Pathologies." *Advances in Neural Information Processing Systems* 34: 15395–15407.

Chandrasekharan, E., S. Jhaver, A. Bruckman, and E. Gilbert. 2022. "Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit." *ACM Transactions on Computer-Human Interaction (TOCHI)* 29 (4): 1–26.

Chandrasekharan, E., U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert. 2017. "You Can't Stay Here: The Efficacy of Reddit's 2015 ban Examined Through Hate Speech." *Proceedings of the ACM on Human-Computer Interaction* 1 (CSCW): 1–22.

Char, D. S., M. D. Abràmoff, and C. Feudtner. 2020. "Identifying Ethical Considerations for Machine Learning Healthcare Applications." *The American Journal of Bioethics* 20 (11): 7–17.

Chollet, F., and J. J. Allaire. 2018. *Deep Learning with R*. Manning Publications Co.

Chouldechova, A., and A. Roth. 2020. "A Snapshot of the Frontiers of Fairness in Machine Learning." *Communications of the ACM* 63 (5): 82–89.

Ciampaglia, G. L., P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini. 2015. "Computational Fact Checking from Knowledge Networks." *PLoS One* 10 (6): e0128193.

Cronk, L. 1999. *That Complex Whole: Culture and the Evolution of Human Behavior*. New York: Routledge.

Cruz, A., G. Rocha, R. S. Silva, and H. L. Cardoso. 2019. "Team Fernando-Pessa at SemEval-2019 Task 4: Back to Basics in Hyperpartisan News Detection." In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

Dahlin, E. 2021. "Mind the gap! On the Future of AI Research." *Humanities and Social Sciences Communications* 8 (1): 1–4.

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Durkheim, É. [1915]1965. *The Elementary Forms of the Religious Life [Trans: Swain JW]*. New York: Free Press.

Eilders, C. 2000. "Media as Political Actors? Issue Focusing and Selective Emphasis in the German Quality Press." *German Politics* 9 (3): 181–206.

Epstein, Z., N. Foppiani, S. Hilgard, S. Sharma, E. Glassman, and D. Rand. 2022, May. "Do Explanations Increase the Effectiveness of AI-Crowd Generated Fake News Warnings?" *Proceedings of the International AAAI Conference on Web and Social Media* 16: 183–193.

Ethayarajh, K., and D. Jurafsky. 2020. "Utility is in the Eye of the User: A Critique of NLP Leaderboards." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4846–4853.

Ghanem, B., S. P. Ponzetto, P. Rosso, and F. Rangel. 2021. Fakeflow: Fake News Detection by Modeling the Flow of Affective Information, *arXiv preprint arXiv:2101.09810*.

Gillespie, T. 2020. "Content Moderation, AI, and the Question of Scale." *Big Data & Society* 7 (2): 2053951720943234.

Grinberg, N., K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer. 2019. "Fake News on Twitter During the 2016 US Presidential Election." *Science* 363 (6425): 374–378.

Grover, A., and J. Leskovec. 2016. "node2vec: Scalable Feature Learning for Networks." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 855–864.

Gruppi, M., B. D. Horne, and S. Adalı. 2021. Tell Me Who Your Friends Are: Using Content Sharing Behavior for News Source Veracity Detection. *arXiv preprint arXiv:2101.10973*.

Gruppi, M., B. D. Horne, and S. Adalı. 2022a. Nela-gt-2021: A Large Multi-Labelled News Dataset for the Study of Misinformation in News Articles. *arXiv preprint arXiv:2203.05659*.

Gruppi, M., B. D. Horne, and S. Adali. 2018. *An Exploration of Unreliable News Classification in Brazil and the US*.

Gruppi, M., P. Smeros, S. Adalı, C. Castillo, and K. Aberer. 2022b. SciLander: Mapping the Scientific News Landscape. arXiv preprint arXiv:2205.07970.

Hassan, N., G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, and M. Tremayne. 2017. "Claimbuster: The First-Ever End-to-End Fact-Checking System." *Proceedings of the VLDB Endowment* 10 (12): 1945–1948.

Heidari, M., S. Zad, P. Hajibabaee, M. Malekzadeh, S. HekmatiAthar, O. Uzuner, and J. H. Jones. 2021. "BERT Model for Fake News Detection Based on Social Bot Activities in the Covid-19 Pandemic." In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 0103–0109). IEEE.

Herkert, J. R. 2005. "Ways of Thinking About and Teaching Ethical Problem Solving: Microethics and Macroethics in Engineering." *Science and Engineering Ethics* 11 (3): 373–385.

Horne, B., and S. Adali. 2017. "This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body,

More Similar to Satire Than Real News." *Proceedings of the International AAAI Conference on Web and Social Media* 11 (1): 759–766.
Horne, B. D., W. Dron, S. Khedr, and S. Adali. 2018. "Assessing the News Landscape: A Multi-Module Toolkit for Evaluating the Credibility of News." In *Companion Proceedings of the The Web Conference 2018*, 235–238.
Horne, B. D., D. Nevo, J. O'Donovan, J. H. Cho, and S. Adalı. 2019. "Rating Reliability and Bias in News Articles: Does AI Assistance Help Everyone?" *Proceedings of the International AAAI Conference on Web and Social Media* 13: 247–256.
Horne, B. D., J. Nørregaard, and S. Adali. 2019. "Robust Fake News Detection Over Time and Attack." *ACM Transactions on Intelligent Systems and Technology (TIST)* 11 (1): 1–23.
Hutchinson, B., N. Rostamzadeh, C. Greer, K. Heller, and V. Prabhakaran. 2022. "Evaluation Gaps in Machine Learning Practice." In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1859–1876.
Huyen, Chip. 2022. *Designing Machine Learning Systems*. Sebastopol, CA: O'Reilly Media.
Iyengar, S., and K. S. Hahn. 2009. "Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use." *Journal of Communication* 59 (1): 19–39.
Jackson, S. 2019. "The Double-Edged Sword of Banning Extremists from Social Media." https://osf.io/preprints/socarxiv/2g7yd/.
Jwa, H., D. Oh, K. Park, J. M. Kang, and H. Lim. 2019. "Exbake: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (Bert)." *Applied Sciences* 9 (19): 4062.
Kaliyar, R. K., A. Goswami, and P. Narang. 2021. "FakeBERT: Fake News Detection in Social Media with a BERT-Based Deep Learning Approach." *Multimedia Tools and Applications* 80 (8): 11765–11788.
Kant, I. 1998[1785]. *Groundwork of the Metaphysics of Morals*. Cambridge: Cambridge University Press.
Katsaros, M., K. Yang, and L. Fratamico. 2022. "Reconsidering Tweets: Intervening During Tweet Creation Decreases Offensive Content." *Proceedings of the International AAAI Conference on Web and Social Media* 16: 477–487.
Kearns, M., and A. Roth. 2020. "Ethical Algorithm Design." *ACM SIGecom Exchanges* 18 (1): 31–36.
Kirkpatrick, K. 2016. "Battling Algorithmic Bias: How Do We Ensure Algorithms Treat Us Fairly." *Communications of the ACM* 59 (10): 16–17.
Kitto, K. L., and B. Sylvester. 2002. "A Multidisciplinary Approach to Teaching Ethical Considerations in Engineering Technology," In *32nd Annual Frontiers in Education, Boston, MA, USA*, pp. S3F–S3F. https://doi.org/10.1109/FIE.2002.1158705.
Koch, B., E. Denton, A. Hanna, and J. G. Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. *arXiv preprint arXiv:2112.01716*.
Kordzadeh, N., and M. Ghasemaghaei. 2022. "Algorithmic Bias: Review, Synthesis, and Future Research Directions." *European Journal of Information Systems* 31 (3): 388–409.
Kuhn, T. S. 2012. *The Structure of Scientific Revolutions* (50th ed.). University of Chicago Press.
Kula, S., M. Choraś, and R. Kozik. 2021. "Application of the Bert-Based Architecture in Fake News Detection." In *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020)* 12: 239–249. Springer International Publishing.
Kumar, S., and N. Shah. 2018. False Information on Web and Social Media: A Survey. *arXiv preprint arXiv:1804.08559*.
Lakkaraju, H., J. Kleinberg, J. Leskovec, J. Ludwig, and S. Mullainathan. 2017. "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284.
Lazer, D. M., M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, and J. L. Zittrain. 2018. "The Science of Fake News." *Science* 359 (6380): 1094–1096.
Lebovitz, S., N. Levina, and H. Lifshitz-Assaf. 2021. "Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What." *MIS Quarterly* 45 (3): 1501–1526.
Lee, N., Z. Liu, and P. Fung. 2019. "Team yeon-zi at Semeval-2019 Task 4: Hyperpartisan News Detection by De-Noising Weakly-Labeled Data." In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 1052–1056.
Lewandowsky, S., U. K. Ecker, and J. Cook. 2017. "Beyond Misinformation: Understanding and Coping With the "Post-Truth" Era." *Journal of Applied Research in Memory and Cognition* 6 (4): 353–369.
Liao, T., R. Taori, I. D. Raji, and L. Schmidt. 2021. "Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning." In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
Martin, K. 2019. "Designing Ethical Algorithms." *MIS Quarterly Executive June*.
Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. 2021. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys (CSUR)* 54 (6): 1–35.
Metaxas, P. T., S. Finn, and E. Mustafaraj. 2015. "Using twittertrails.com to Investigate Rumor Propagation." In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, 69–72.
Mill, John Stuart. 2002. *The Basic Writings of John Stuart Mill: On Liberty, the Subjection of Women, and Utilitarianism*. New York: Modern Library.
Molewijk, B., A. M. Stiggelbout, W. Otten, et al. 2004. "Empirical Data and Moral Theory. A Plea for Integrated Empirical Ethics." *Medicine, Health Care, and Philosophy* 7: 55–69.
Mosallanezhad, A., M. Karami, K. Shu, M. V. Mancenido, and H. Liu. 2022. "Domain Adaptive Fake News Detection Via Reinforcement Learning." In *Proceedings of the ACM Web Conference 2022*, 3632–3640.
Newman, T. P., E. C. Nisbet, and M. C. Nisbet. 2018. "Climate Change, Cultural Cognition, and Media Effects: Worldviews Drive News Selectivity, Biased Processing, and Polarized Attitudes." *Public Understanding of Science* 27 (8): 985–1002.
Nickerson, R. S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2: 175–220.
Nørregaard, J., B. D. Horne. 2019. "NELA-GT-2018: A Large Multi-labelled News Dataset for the Study of

Misinformation in News Articles." *Proceedings of the International AAAI Conference on Web and Social Media* 13: 630–638.
Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–453.
Patricia Aires, V., G. Nakamura, and E. Nakamura. 2019. "A Link-Based Approach to Detect Media Bias in News Websites." In *Companion Proceedings of The 2019 World Wide Web Conference*, 742–745.
Potthast, M., J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein. 2017. A Stylometric Inquiry Into Hyperpartisan and Fake News. *arXiv preprint arXiv:1702.05638*.
Prabhakaran, V., R. Qadri, and B. Hutchinson. 2022. Cultural Incongruencies in Artificial Intelligence. *arXiv preprint arXiv:2211.13069*.
Reis, J. C., A. Correia, F. Murai, A. Veloso, and F. Benevenuto. 2019. "Supervised Learning for Fake News Detection." *IEEE Intelligent Systems* 34 (2): 76–81.
Resnick, P., S. Carton, S. Park, Y. Shen, and N. Zeffer. 2014. "Rumorlens: A System for Analyzing the Impact of Rumors and Corrections in Social Media." In *Proc. Computational Journalism Conference*, 5(7).
Ribeiro, M. H., J. Cheng, and R. West. 2022. "Post Approvals in Online Communities." *Proceedings of the International AAAI Conference on Web and Social Media* 16: 335–346.
Riley, P. 2019. "Three Pitfalls to Avoid in Machine Learning." *Nature* 572 (7767): 27–29.
Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–1082.
Robinson, S. 2020. "*Trust, Transparency, and Openness: How Inclusion of Cultural Values Shapes Nordic National Public Policy Strategies for Artificial Intelligence (AI).*" Technol. Soc. 63:1014–1021.
Rodriguez, P., J. Barrow, A. M. Hoyle, J. P. Lalor, R. Jia, and J. Boyd-Graber. 2021. "Evaluation Examples Are Not Equally Informative: How Should That Change NLP Leaderboards?." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4486–4503.
Ruchansky, N., S. Seo, and Y. Liu. 2017. "Csi: A Hybrid Deep Model for Fake News Detection." In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 797–806.
Sculley, D., J. Snoek, A. Wiltschko, and A. Rahimi. 2018. "Winner's Curse? On Pace, Progress, and Empirical Rigor." In *Proceedings of ICLR 2018*.
Shabel, L. 2017. "Kant's Mathematical Principles of Pure Understanding." In *Kant's Critique of Pure Reason: A Critical Guide*, edited by J. O'Shea, 163–183. Cambridge, UK: Cambridge University Press.
Shu, K., H. R. Bernard, and H. Liu. 2019a. "Studying Fake News Via Network Analysis: Detection and Mitigation." *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, 43–65.
Shu, K., D. Mahudeswaran, S. Wang, and H. Liu. 2020. "Hierarchical Propagation Networks for Fake News Detection: Investigation and Exploitation." In *Proceedings of the International AAAI Conference on Web and Social Media*, 14:626–637.
Shu, K., A. Sliva, S. Wang, J. Tang, and H. Liu. 2017. "Fake News Detection on Social Media: A Data Mining Perspective." *ACM SIGKDD Explorations Newsletter* 19 (1): 22–36.
Shu, K., S. Wang, and H. Liu. 2018. "Understanding User Profiles on Social Media for Fake News Detection." In *2018 IEEE conference on multimedia information processing and retrieval (MIPR),* 430–435.
Shu, K., S. Wang, and H. Liu. 2019b. "Beyond News Contents: The Role of Social Context for Fake News Detection." In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 312–320.
Singhal, S., R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. I. Satoh. 2019. "Spotfake: A Multi-Modal Framework for Fake News Detection." In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 39–47.
Stahl, B. C. 2006. "On the Difference or Equality of Information, Misinformation, and Disinformation: A Critical Research Perspective." *Informing Science* 9: 83.
Swire-Thompson, B., J. DeGutis, and D. Lazer. 2020. "Searching for the Backfire Effect: Measurement and Design Considerations." *Journal of Applied Research in Memory and Cognition* 9 (3): 286–299.
Swire-Thompson, B., N. Miklaucic, J. P. Wihbey, D. Lazer, and J. DeGutis. 2022. "The Backfire Effect After Correcting Misinformation is Strongly Associated with Reliability." *Journal of Experimental Psychology: General* 151 (7): 1655.
Szczepański, M., M. Pawlicki, R. Kozik, and M. Choraś. 2021. "New Explainability Method for BERT-Based Model in Fake News Detection." *Scientific Reports* 11 (1): 23705.
Sztandar-Sztanderska, K., and M. Zielenska. 2018. "Changing Social Citizenship Through Information Technology." *Social Work & Society* 16 (2): 1–13.
Trujillo, M., M. Gruppi, C. Buntain, and B. D. Horne. 2020. "What is Bitchute? Characterizing the "Free Speech" Alternative to YouTube." In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, 139–140.
Van der Linden, S., A. Leiserowitz, S. Rosenthal, and E. Maibach. 2017. "Inoculating the Public Against Misinformation About Climate Change." *Global Challenges* 1 (2): 1600008.
Varnum, M. E., and I. Grossmann. 2017. "Cultural Change: The how and the why." *Perspectives on Psychological Science* 12 (6): 956–972.
Vosoughi, S., D. Roy, and S. Aral. 2018. "The Spread of True and False News Online." *science* 359 (6380): 1146–1151.
Wadden, J. J. 2022. "Defining the Undefinable: The Black box Problem in Healthcare Artificial Intelligence." *Journal of Medical Ethics* 48 (10): 764–768.
Wang, W. Y. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. *arXiv preprint arXiv:1705.00648*.
Wardle, C., and H. Derakhshan. 2017. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking, 27:1-107*. Strasbourg: Council of Europe.
Yang, K. C., and F. Menczer. 2023. Large Language Models Can Rate News Outlet Credibility. arXiv preprint arXiv:2304.00228.

Yang, S., K. Shu, S. Wang, R. Gu, F. Wu, and H. Liu. 2019. "Unsupervised Fake News Detection on Social Media: A Generative Approach." *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (1): 5644–5651.

Yaser, S. A., M. Magdon-Ismail, and H. T. Lin. 2012. *Learning From Data: A Short Course*.

Zagzebski, Linda. 1997. "Virtue in Ethics and Epistemology." *Proceedings of the American Catholic Philosophical Association* 71: 1–17.

Zannettou, S. 2021. ""I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter." *Proceedings of the International AAAI Conference on Web and Social Media* 15: 865–876.

Zannettou, S., M. Sirivianos, J. Blackburn, and N. Kourtellis. 2019. "The web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans." *Journal of Data and Information Quality (JDIQ)* 11 (3): 1–37.

Zhang, J. M., M. Harman, L. Ma, and Y. Liu. 2020. "Machine Learning Testing: Survey, Landscapes and Horizons." *IEEE Transactions on Software Engineering* 48 (1): 1–36.

Züger, T., and H. Asghari. 2023. "AI for the Public. How Public Interest Theory Shifts the Discourse on AI." *AI & Soc* 38: 815–828.