# Ethical Considerations in Data-Driven Fake News Detection


**Deepak P**

**Abstract** Data-driven and AI-based detection of fake news has seen much recent interest. The focus of research on data-driven fake news detection has been on developing novel and effective machine learning pipelines. The field has flourished with the rapid advances in deep learning methodologies and the availability of several labelled datasets to benchmark methods. While treating fake news detection as yet another data analytics problem, there has been little work on analyzing the ethical and normative considerations within such a task. This work, in a first-of-its-kind effort, analyzes ethical and normative considerations in using data-driven automation for fake news detection. We first consider the ethical dimensions of importance within the task context, followed by a detailed discussion on adhering to fairness and democratic values while combating fake news through data-driven AI-based automation. Throughout this chapter, we place emphasis on acknowledging the nuances of the digital media domain and also attempt to outline technologically grounded recommendations on how fake news detection algorithms could evolve while preserving and deepening democratic values within society.

**Keywords** Ethics · Fairness · Fake news detection · Data science

## 1 Introduction

Data-driven fake news detection involves the usage of machine learning and data analytics methods in order to combat fake news. It is still early days in this discipline, and thus algorithms in this space have largely explored supervised methods for the task, with some limited work on unsupervised fake news detection. Active learning, transfer learning, and reinforcement learning are not yet popular for fake news detection. With the growing ecosystem of fake news and a widespread recognition of the pervasiveness of fake news or disinformation through buzzwords like *post-truth*, fake news is arguably something we would need to live with in the



Deepak P et al., *Data Science for Fake News*, The Information Retrieval Series 42, https://doi.org/10.1007/978-3-030-62696-9_10

long run. *Collins Dictionary* chose *fake news* as the word of the year in 2017,[1] in the aftermath of the 2016 US Presidential Elections during which the phrase was used heavily. Given these trends, one could envisage fake news detection as being embedded by default in various information delivery platforms in the future, much in the same way that spam detection has become a default feature offered by most email service providers. As this chapter is being authored, Microsoft has started including its *NewsGuard* plugin, which rates website credibility, in mobile versions of its Edge browser[2] as a feature turned on by default; Fig. 1 shows the NewsGuard plugin in action.

In this chapter, we consider the ethical aspects of data-driven fake news detection (DFND). As a first work in this topic, we endeavor to consider a broad set of ethical dimensions in DFND. We place emphasis on ensuring that this chapter is understandable for a broad audience much beyond technologists working in the area, providing abundant context wherever necessary. We outline several ethical dimensions that are pertinent for DFND in Sect. 2. This is then followed by a discussion of fairness in DFND in Sect. 3. We consider aspects around the uptake of DFND in Sect. 4, especially from the perspective of how democratic values could be presented during the course of such uptake; in this section, we also endeavor to provide some concrete recommendations that could help guide AI approaches to fake news detection. We then conclude the chapter in Sect. 5.

## 2 Ethical Dimensions of DFND

The ethical considerations in DFND fall under the broad umbrella of ethical considerations of any data-driven optimization task but are confounded greatly by the societal importance of the task. Thus, the domain of fake news poses some unique ethical considerations in that it operates in a domain where it seeks to make judgments on news and could thus influence opinions and substantive decisions made by humans. In order to illustrate the contrast with other domains, consider product recommendation, the task of determining whether a user may like a product or not. This task is relatively benign in moral terms in that a bad prediction may only result in users receiving bad product suggestions. For example, a chocolate ice cream lover may be sent offers pertaining to vanilla ice creams due to an inaccurate decision, or a beer lover could be sent wine recommendations. While these are evidently problematic, their effects are limited to creating user frustration and/or leading them into bad choices but (arguably) have limited impact beyond the

[1] https://www.independent.co.uk/news/uk/home-news/fake-news-word-of-the-year-2017-collins-dictionary-donald-trump-kellyanne-conway-antifa-corbynmania-a8032751.html.

[2] https://www.cnet.com/news/microsofts-edge-browser-warns-you-about-fake-news/.

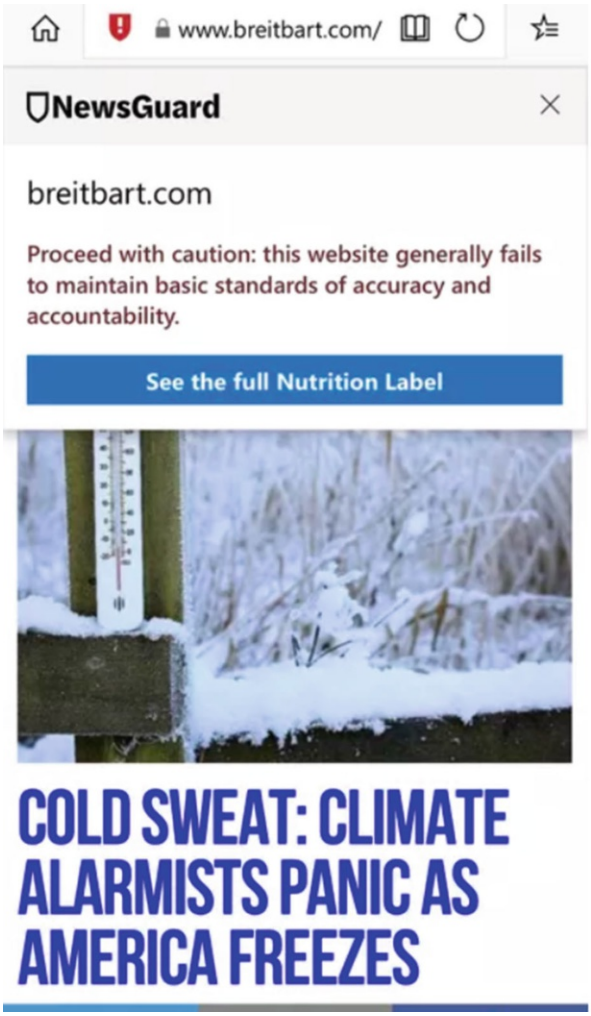


**Fig. 1** Edge NewsGuard Plugin displaying a warning

purchase and consumption of the product. On the other hand, fake news on climate change being labelled as *non-fake* has serious ramifications. It could sway individual users' and public opinion away from green policies and could be harmful to society as a whole. Similarly, xenophobic fake news has been increasingly used as a tool by certain political parties to sway public opinion toward themselves.

We analyze the space of ethical considerations of DFND across three dimensions, which are briefly outlined herein:

– **Mismatch of Values:** A core ethical consideration in the context of data and AI technologies whose growth has been fuelled by automation efficiency and other values of the market is the tension between the values embedded in them and the values held in society. This includes the tension between *accuracy* and *fairness*

as well as that between *convenience* and *dignity* and others considered in various contexts [21]. This conflict is also relatable to positions in the political spectrum in reasonably unambiguous ways.

- **Nature of Data-Driven Learning:** An important ethical consideration comes from the nature of data-driven algorithms themselves. Data-driven algorithms look to build statistical models from the past (past encoded in historical datasets used for training) and attempt to use such models for the future. While this is done explicitly during the training process in the case of supervised learning, assumptions based on the past are implicitly encoded within the design of both supervised and unsupervised algorithms. This involves an implicit assumption of a static nature of the high-level data and application scenario, which causes ethical ramifications in the domain of fake news.
- **Domain Properties:** There are certain properties of the domain that spawn ethical and normative considerations. As a simple example, unlike ad recommendations where the same ad could be relevant for a user and irrelevant for another, a news article judged to be fake needs to be judged fake for all users. Further, certain other inconsistencies may be inadmissible. As an example, the veracity decision should not depend on who is quoted in the article; in other words, fake news should be judged as fake regardless of who is quoted as relaying it.

We will delve into such ethical considerations in detail in the following subsections. We do not claim that this covers the full spectrum of ethical considerations but do hope to cover many important ones.

## *2.1 Mismatch of Values*

We will now consider ethical ramifications from the mismatch of values for which ML algorithms are designed to optimize and those expected in the application domains such as fake news detection. We will consider the historical context of ML, and how things have changed from thereon, and outline various ethical facets of the value mismatch.

**Historical Context of ML** It helps to consider a historical perspective of machine learning in order to understand the context of the ethical considerations that emanate from the *mismatch of values*. The initial efforts of machine learning were targeted toward automating tasks that were inappropriate or difficult to be automated by means of rule-based methodologies. As an oft-quoted example, consider the case of handwriting recognition. It is extremely hard, if not impossible, to come up with a set of rules that would fit together to form a system to effectively recognize handwritten text. Tasks such as handwriting recognition are, by nature, tasks that humans are quite good at but often perceive as quite mundane. Data entry tasks that involve handwriting recognition were considered within the lowest rung of IT-enabled jobs in terms of skills required. Thus, machine learning, in its early days, aimed to automate such mundane tasks using abundant historical traces of

human performance over such tasks to learn statistical models that would help replace or reduce manual labor spent on the task. Most early advances in machine learning were around tasks of a similar nature such as image recognition, text to speech translation, and automation of search. All of these target the optimization of mundane tasks for which the natural metric of success is *amount of labor automated*. Such is the case with another application realm for machine learning, that of robotics, where physical labor was sought to be automated. With optimization of manual labor being a priority for businesses who wanted to improve their competitive advantage in an emerging IT-oriented marketplace, investment in machine learning was aligned with market priorities. Cumulative metrics such as *precision*, *recall*, and *accuracy* were the natural targets for optimization, since they are easily translatable into automated cumulative labor. In semi-automated machine learning pipelines where humans were always available to correct errors, such as a supervisor who would gloss over handwriting recognition outputs to correct any apparent errors, the quantum of manual effort is the obvious area to be minimized. Such settings, and their more automated counterparts, did not offer any incentive to consider the *distribution of errors*. As a hypothetical example, a system which always misidentified a particular rare word, say a complex one such as *xylophone*, would be acceptable if that misidentification helped the model move toward such directions that ensure correct identification of a number of other common words. Within the historical context of machine learning envisaged as a minor and passive player that seeks to automate a set of mundane tasks within a sophisticated ecosystem, such market-driven and efficiency-oriented considerations being the sole or primary consideration made natural sense. These automation-oriented metrics, in the landscape of political philosophy, align with the schools of utilitarianism [19]. This school of thought seeks to maximize cumulative good, which translates well into automation being the target of maximization in the case of ML.

**Current ML Applications** Of late, one may observe that two major changes have taken place. First, the realization of the power of data-driven learning began to accelerate the uptake of machine learning quite dramatically. Machine learning algorithms started to become *major players* (as against minor ones) in an increasingly IT-enabled ecosystem and consequently started playing a more *active role* than a merely passive one. Second, machine learning started to be applied for tasks much beyond the original limited remit of *mundane tasks* that are worthy of automation. As Narayanan [15] opines, ML has been moving from the domain of *automating perception* (e.g., tasks such as handwriting recognition and face recognition) to domains of *automating judgment* (e.g., spam detection, fake news detection) and *predicting social outcomes* (e.g., predictive policing, predicting criminality from a face! [8]). Narayanan argues that the task of predicting social outcomes can be regarded as fundamentally dubious from an ethical perspective. The interplay between the first factor (pervasive use of ML) and the second factor (usage for predicting social outcomes) leads to serious issues that may not be apparent when considering them separately. As an example from an ML use case from predictive

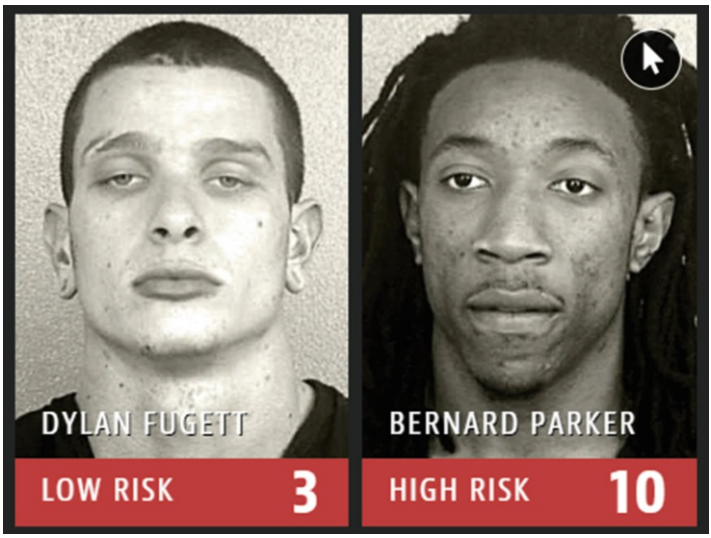


**Fig. 2** The risk scores from COMPAS for two individuals detained for drug possession, illustrating racial bias (image Source: ProPublica)

policing, an initial preference for labelling minority areas as crime prone can be imbibed by an algorithm that aims to identify areas for higher surveillance. Crimes are caught only when *committed* as well as *observed*; higher surveillance in minority areas increases the observation rate of committed crimes, reinforcing the bias. Thus, not only does ML enable institutionalizing bias inherent within historical data, it creates even more biased data for the next generation of tools to work on, compounding the problem. Thus, machine learning algorithms are today employed in making decisions that significantly affect human lives. In what has become a widely cited example of bias, COMPAS, a software tool to predict recidivism in the USA, has been widely criticized for being biased against blacks; Fig. 2 shows the *risk scores* assigned by COMPAS to two individuals detained on account of possession of drugs.

---

*An evolutionary perspective predicts that most diets and fitness programs will fail, as they do, because we still do not know how to counter once-adaptive primal instincts to eat donuts and take the elevator.*

—Daniel E. Lieberman in "The Story of the Human Body: Evolution, Health, and Disease" [13]

---

**The Facets of the Mismatch** While the historical context of automation of mundane tasks made cumulative efficiency-oriented market-aligned metrics the natural ones to optimize for, the new application scenarios make them least suited due to their conflicts with the values of society. ML algorithms that have historically

been advanced along a certain direction (guided by cumulative efficiency-oriented metrics) now need to be steered in a different direction! Such *mismatches* are hardly unique to machine learning; the most studied mismatch is that of *evolutionary mismatch* [14], which refers to evolved traits that were once advantageous but became maladaptive due to changes in the environment. An oft-quoted example is that of human diet, where humans, having evolved for long durations in the African savanna, developed a penchant for rare foods that contain both sugar and fat, as Daniel Lieberman states in the quote cited above. This evolutionary liking encourages us to seek out foods high in fat and sugar, which the market has overtly exploited through abundant placement in supermarket shelves, leading to a pervasive obesity problem in the population. While the analogy does not go the whole way, since ML algorithms are different in being actively designed by humans rather than evolving through natural selection, it does suggest that adapting to the needs of the new tasks is likely to require radical reimagination as opposed to patchwork fixes. The mismatch of values between those *from markets* and those *in society* has several facets, some of which we examine below, within the context of fake news detection:

- **Utility vs. Fairness:** While fake news algorithms should rightly aim to develop the capability to debunk as much fake news as possible (i.e., high utility in terms of fraction of fake news debunked), this should not come at the cost of an asymmetry along facets that matter. For example, even if fake news about *tapeworm* is only 2% of medical fake news, a method that is totally unable to capture that space of fake news would not be acceptable. In other words, the cumulative accuracy/inaccuracy should not have a high distributional skew along facets that are reasonably important.
- **Problematic Features:** Typically, ML algorithms are designed by making use of all features that can potentially tell something about the target variable, since it would help the ML algorithm achieve better accuracy. Thus, if a particular user handle, $U$, is largely used to share fake news, an ML algorithm may learn that pattern. This could, for example, be through a high value of conditional probability $P(fake|U)$ or more sophisticated mechanisms. However, such a feature could be problematic to use, since it undermines the user's (or, for that matter, any human's) ability to evolve, and if such estimates are used widely and the user's posts are blocked more often than not, it disenfranchises the user's voice in the media. It may be argued that something that has a systematic bias against a particular user cannot be construed as being part of a *fair process*. A similar argument could be used, though less compellingly, against using news source IDs as a feature. It may also be noted that simply not using a particular feature may not be enough, since there could be other proxy features. For example, a user may be identifiable through a distinctive language style, and thus the user's correlation with fake news may be learnt indirectly by an ML algorithm.
- **Responsibility and Accountability:** In applications of ML which fall under the *automating perception* category (in Narayanan's categorization [15]), it

was plausible to make an argument that *the further we go, the better*. In other words, it was possible to argue that some amount of automation is better than no automation, and more is better than less. However, when it comes to tasks such as fake news detection, the fact that ML is being used or claimed to be used in this regard implicitly involves much more responsibility. This means that deepening of automation may need to be held off until there is capacity to shoulder the responsibility that comes with such higher levels of automation. There are at least two fronts of responsibility and accountability that come from functioning in a democratic society:

- To the media sources whose news stories are being labelled as fake or non-fake
- To the user who is expected to consume the decision made by the algorithm

It is still an open question as to how these responsibilities may be fulfilled. One possibility could be that a trail or explanation is generated to support the decision, which can be made public, so as to be challenged or debated upon. Then again, should these be subject to legal regulations? If a legal framework needs to be instituted, it would require that the process of ensuring compliance with the legal regulations be laid out clearly. It could also be argued that such enforcements should not be made by legal frameworks but through voluntary compliance with ethical standards developed in the community.

## *2.2 Nature of Data-Driven Learning*

---

*Big Data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that's something only humans can provide.*

—Cathy O' Neil in "Weapons of Math Destruction" [16]

---

We now consider ethical issues that emanate from the very nature of ML or data-driven learning. The broad task in data-driven machine learning is to make use of historical/past data (in conjunction with several other constraints coming from an understanding of the domain) in order to make meaningful decisions about the future. Any perspective that is historically rooted would pooh-pooh a proposal that aims to make an assumption that the past is predictive of the future *when it comes to making decisions on substantive societal issues*. So, how did we come about to even attempting to use ML for such tasks? The answer once again lies in the historical context in which ML developed.

Following on from the narrative in the previous section, we can see that the simple hypothesis of *past predicts the future*—more technically, that training and testing data come from the same distribution—works exceedingly well for tasks such as speech recognition, or characterizing supermarket purchase patterns, especially in the short term. Handwriting styles remain quite static for an individual, and the nature of errors made by ML algorithm handwriting recognition does

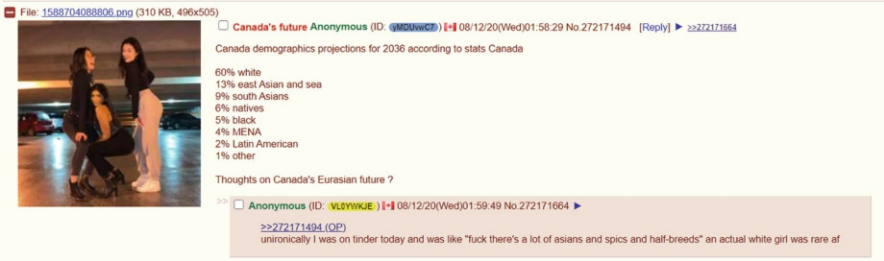


**Fig. 3** A post from the Discussion Board */pol/* from *4chan*, a discussion board often noted for extremist political ideologies

not influence how the person would change his or her handwriting. Similarly, people generally have some amount of periodicity in purchasing regularly used FMCG products, and stores organizing products based on purchase patterns, while enhancing convenience, are not likely to affect consumer purchasing behavior. However, when one considers other domains of activity and the long term, people do evolve substantially. Peddlers of fake news work in a highly dynamic ecosystem of social media platforms, where certain features are more useful than others for propagating fake news. For example, anonymous posting functionalities provided by social networks like *4chan* have been regarded as being exploited heavily by agents that drive fake news.[3] Figure 3 shows a post from *4chan*'s */pol/* discussion board which has been noticed for extremist political ideology as well as alternative facts. WhatsApp recently restricted its forwarding functionality in view of fake news.[4] Such measures lead to a gamification between fake news peddlers and social media platforms, in turn leading to an ever-changing character of fake news, limiting the ability of using historical data in predicting the future. Viewed from another perspective, naively learning from historical data without accounting for the dynamics of the space would lead to techniques that would be biased in being able to discover certain kinds of fake news more than others.

The dynamic environment that exists in the misinformation space escalates in volatility even further with the presence of ML-based fake news detection as an active player. When certain techniques for fake news detection gain prominence and get widely applied, incentives to devise workarounds also emerge along with it. The resultant gamification between fake news detection techniques and fake news itself would lead to a perpetual race by each party to stay one step ahead of the curve. In such a scenario, the nature of fake news detection techniques will also decide the future nature of fake news, and vice versa. This could result in fake news detection mechanisms employing highly complex decision surfaces to stay current

[3] https://news.sky.com/story/research-examines-fake-news-hate-speech-and-4chan-10910915.

[4] https://www.theverge.com/2020/4/7/21211371/whatsapp-message-forwarding-limits-misinformation-coronavirus-india.

and usable. The change in the behavior of fake news detection methods need not be due to conscious engineering by data scientists. The same algorithms when fed with newer labelled data encoding the changes in character of fake news will itself result in changes in the nature of the models built by the same learning methods. In a way, the same ML working as a meta-model using inductive learning will produce different models in response to different labelled datasets.

The highly volatile landscape with multiple actors trying to outpace one another is not quite new. It exists in the case of other domains, a very relatable one being *antivirus* software and, to a lesser extent, *spam detection* software. The makers of antivirus software and the makers of viruses are always in a relationship similar to that between fake news debunking software and fake news peddlers. The difference is that while virus makers are keen on finding new ways to squeeze *self-replicating code* into machines, fake news creators operate in ways to *sway the user's thoughts* in directions that suit their political or economic interests. While the former may be argued to be *morally neutral*, the latter definitely is not so. The intent of fake news differs in ways in which computer viruses do not; political fake news that is "useful" for one party would be "harmful" for another. Since fake news operates in the space of *swaying user opinions and thoughts*, one which has plentiful moral dimensions, care needs to be accorded to how detection algorithms are built.

**Staying Updated in a Volatile Ecosystem** An ML-based fake news detection method that is out of sync with the configuration of the ecosystem over which it would be used for fake news detection could result in a plethora of inaccurate decisions. Reactively adjusting to discovered errors may not be sufficient. This is so since some errors may never be discovered, or are less likely to be discovered; for example, news labelled as fake by a detection method may be hidden from view (depending on how the method is embedded within a software tool), and thus there may not be an opportunity to identify such false positives. Consistently making erroneous decisions that curtail the propagation and visibility of certain opinions can be argued to stand against the spirit of democracy and compromise *reasonable pluralism* [6] in public discourse; this aspect makes this issue distinct from analogous scenarios within antivirus and spam detection software. Continuously procuring a current set of labelled data followed by extensive benchmarking and method refinement may not be feasible due to resource and economic considerations. Nevertheless, a continuously updated conceptual picture of the ecosystem within which the technique would be embedded needs to be maintained, and the technique needs to be periodically contrasted against it in order to ensure that it is current. In particular, the ML method may need to be refined in two distinct dimensions to remain updated: by varying the training data and by varying the method. We consider important questions in this space, the answers to which may point to directions in which the techniques should be refined.

– **Training Data Curation:** The training data, in the case of supervised methods, determines the capabilities that will be infused into the fake news detection model that is eventually learnt. This makes curation of training data an important

consideration in ensuring that the fake news detection technique stays current. This involves aspects such as the following:

- *How old can the training data be?* Very old training data may be inappropriate to use since they may be obsolete artifacts from an ecosystem that has substantively changed.
- *What is the relevance of training data elements?* Even temporally recent training data elements may be of limited relevance if they are associated with aspects of a media ecosystem that no longer exists. For example, one could argue that a social media post that is sparse in content and rich in emojis may be of limited relevance if that was soon followed by a radical change in the affordances with respect to emojis on the social media platform where it is situated.

– **Technique Design:** Every ML method, implicitly or explicitly, makes use of some assumptions about the domain in order to carry out the learning process. Some of these assumptions may be violated with changes in the media ecosystem that happen due to ML or non-ML actors as outlined above. Within unsupervised learning methods that do not have the luxury of being guided by training data, technique design considerations are more central and worthy of more attention. As an example, a *truth discovery* approach [22] makes an implicit assumption that fake news is represented on a minority of websites and that fake narratives diverge from facts in different directions. The presence of a widespread orchestrated fake news campaign could easily upturn such an assumption and lead the technique to discovering fake news as real and vice versa. Similarly, behavioral heuristics such as assumptions on synchronous user activity employed by recent methods (e.g., [7]) could also be invalidated by novel strategies by fake news peddlers.

## *2.3 Domain Properties*

Fake news detection operates in the space of media, often referred to as the *fourth estate*, a space where actors have significant indirect influence in the political ecosystem. Further, the nature of the media domain entails some unique considerations for AI interventions within it. We outline some such unique aspects below:

– **Veracity Decisions as Impersonal and Universal:** This is an era of personalization, where ML algorithms routinely make use of user profiles to tailor decisions to them. We compared fake news detection with spam detection many times over, and within spam there is an element of personalization that could be legitimately brought in by making use of the inputs from the user on what is spam for them and what is not. Indeed, most ML-based personalized algorithms operate at two levels: one that makes use of cross-user data to learn general trends across a large

dataset and another that makes use of user-specific data to learn specific likings of the user. The decision for a user, such as whether an email is spam, is one that blends both these factors. Thus, an email that goes into the spam folder for a user may legitimately need to land in the inbox for another. However, such personalization is inherently incompatible with the task of fake news detection, since there is no reason why a news is *fake* for one and *legitimate* for another. This aspect needs to be seen normatively, rather than in terms of utility. For example, *fake news* on vaccines and autism may be comforting for an *anti-vax* activist, and thus personalization that does not flag the news as fake may be better for improving user satisfaction for him/her, the utilitarian metric that most such methods aim to optimize. Despite such factors, the veracity decision needs to be consistent across users, from a normative standpoint.

- **Decision Timeliness, Reversals, and Accountability:** The emerging understanding of fake news involves a finding that exposure is hard to correct [18]; in other words, a person exposed to a news is still influenced by it long after it is exposed as fake to the same person. A news delivery system which claims to have a fake news detection functionality thus needs to ensure timely decisions to reduce exposure to fake news, in view of the accountability considerations discussed in an earlier section. It may also be argued that there is value in deferring dissemination of news articles until they are verified, especially if the fake news detection is implemented on a news delivery platform such as Google News. If that is not done (and it may be infeasible to do so in cases where the fake news identification is embedded within a browser plugin, where the user acts independently of the service), it may be argued that the service may be considered accountable to those who read a news in their system which was later labelled as fake. Does the system have an implicit obligation to proactively inform such users about the finding of fakeness? ML systems make decisions on the basis of data. As new data emerges, decisions may have to be reversed, or the confidence in a particular decision may deteriorate to an ambiguous range. It is interesting to analyze, from the perspective of accountability, as to how systems should handle such decision reversals. A somewhat similar case exists in online media where it is considered a good practice to make all edits to a published article public. In any case, there is a higher degree of accountability toward users who viewed an earlier decision that was reversed, as compared to somebody who viewed the article prior to any decision from the fake news detection method. It may also be seen here that these dimensions of accountability around decision timeliness and reversals do not apply to the earlier generation of tasks such as handwriting recognition (at least, not anywhere close to the same extent).
- **Veracity of Reporting or Reported Information:** Consider a case where a famous person, say $X$, makes a verifiably fake claim, such as *turmeric water can cure COVID-19*. What would be the veracity label attached to a news article, or a tweet such as that shown in Fig. 4, that carries the statement: *X says that turmeric water can cure COVID-19?* There are arguments on two sides. First, that the news piece is *non-fake* since $X$ did actually make the claim. Second, that the news piece is *fake* since it contains a verifiably false claim. By treating fake


**Fig. 4** A tweet reports that a famous person claimed a COVID-19 cure. Do we verify whether the claim was correct or whether the reporting was indeed factual? These choices lead to different veracity decisions

news detection as a data science problem, such important nuances could easily be brushed under the carpet by relegating them to the ways in which they are labelled, which in turn may depend on how individual labellers think about them. However, it is important to consider whether fake news detection should restrict itself to superficial verification (e.g., whether the statement reported was actually made) and whether it needs to go a level deeper (i.e., whether the statement is actually true). This could lead to different kinds of fake news detection systems.

## 3 Fairness and DFND

We now consider aspects of fairness and how they apply to the task of DFND. Fairness is used variously and is interpreted as related to a number of other concepts such as *equity*, *impartiality*, *unbiased*, and *equality*. Fairness has a long tradition within philosophy over the centuries, and the most widely accepted usage today could be in the context of *justice as fairness*, Rawls' pioneering work [17] in 1971.

While a broad discussion of fairness is well beyond the scope of this work, we will consider fairness in the way it has been used in machine learning literature, fair ML being a very active area since an early work [5] in 2012.

**Streams of Fairness in Machine Learning** Fairness in machine learning has been studied under two distinct streams: *individual fairness* and *group fairness*. While this distinction has come under recent criticism [3], we will use it as it provides a conceptual distinction between routes for deepening fairness. Individual fairness is commonly interpreted as being related to application of *fair procedure*, in that the task is done without partiality to the individual and in full sincerity to the aspects that matter to the task. A fair job selection process should thus only make use of attributes or features of a candidate that matter to the job and nothing else. In most analytics tasks, this would mean that *similar objects get assigned similar outcomes* and that similarity is assessed in a task-relevant manner. While all of this should come across as natural, what it keeps out of scope is important to analyze. It does not consider the *historical context* of data and interprets *data as given*. Thus, an individually fair or procedurally correct method discards any historical context of entrenched oppression that has caused some ethnicities to be disadvantaged with respect to access to quality education or if the metrics that measure *future productivity* in the job are set up in a way that is advantageous to certain ethnicities. This means that a method that agrees to tenets of individual fairness could produce *unequal outcomes* on dimensions such as ethnicity or other dimensions such as gender within which historical asymmetries exist. On the other hand, *group fairness* algorithms interpret fairness as a property of outcomes. It usually works by designating some attributes as *sensitive*; these are typically attributes that an individual usually does not have much role in determining for herself, or on whom asymmetries in societies usually function. Thus, these could include *gender*, *ethnicity*, *nationality*, and *religion*. Group fair algorithms try to ensure that parity is maintained across such specified sensitive dimensions. For example, if blacks have a one-seventh representation in a population, as is roughly the case in the USA, group fairness would be violated when the proportion of blacks among successful hires deviates much from one-seventh. These constraints are enshrined, though not to the fullest extent, within legal provisions such as the *Uniform Guidelines on Employee Selection Procedures (1978)* [2] in the USA, and provisions for affirmative action by way of quotas (commonly called *reservations*) within the Indian Constitution, viz., Article 15(4).[5]

**Fairness and Impersonal Data** The above notions of fairness are well motivated when making decisions about human beings based on their data. Indeed, the notion of *equity* and *equality* is most supported within a society when it comes to treatment of individuals. This would also naturally extend to cases where certain other attributes are correlated with individuals' *sensitive attributes*. As an example, we may argue that predictive policing methods should enforce some

[5] https://indiankanoon.org/doc/251667/.

kind of parity between minority and non-minority neighborhoods to ensure that societal anti-minority stereotypes are not reinforced by heavily policing minority neighborhoods. Such fairness arguments may also be extended toward geographical regions, where we may expect a *decent* level of public infrastructure across regions. For example, we may want to ensure that road works are not unduly delayed in rural neighborhoods even if the roads are less heavily used there as compared to urban neighborhoods. These issues are particularly of concern in cases where crowdsourcing is used to collect reports on issues such as *report a pothole* services that are being deployed[6] by governments worldwide. Solely relying on such IT-enabled crowdsourcing mechanisms could reinforce existing asymmetries. Rural roads are likely both less busy and residents may be less tech-savvy, both of which could cause underreporting of issues from rural localities. This may be seen as a notion of group fairness when treating *geographical region* as a sensitive attribute. Thus, the applications of principles of fairness could extend beyond personal data and could be carefully and meaningfully extended to data that does not pertain to human beings.

**Fairness and Fake News Detection** How would we go about thinking about the usage of fairness principles and their applications in the task of fake news detection? One possibility is to first consider violations of fairness we would necessarily want to avoid. We discuss some examples here:

- **Political Alignment:** Consider an example where a political party enters the fake news detection business and provides a plugin that explicitly states that *it debunks fake news from its political opponents*. In certain other cases, the political alignment may be less explicit than this but may serve a similar function. Would we want to permit such a fake news detection method even if it truthfully admits the bias? Such a tool may work either by keeping news sources that it favors completely out of the detection remit or by ensuring they are labelled as non-fake through other means. It may be argued that such tools reinforce the echo chamber effects that personalized news is often criticized for [20]. Due to such reasons, such politically aligned fake news detection that is unfair in being biased toward particular political positions may be considered undesirable. Should we then consider political leaning as a *sensitive attribute* and shoot for group fairness? We will consider such options soon.
- **Different Standards:** We may also want to avoid fake news detection tools that apply different standards to different parts of the news domain. Such different standards could emerge from seemingly legitimate reasons. As an example, a fake news detection engine may decide that news from a particular country may be fact-checked against authoritative sources within that country, as a maxim of procedural fairness. This would entail that disparities in authoritativeness between reference sources across various nations would naturally manifest as different standards. As an example, the same news could be assigned different

[6] https://www.nidirect.gov.uk/services/report-pothole.

veracity labels based on which country it stems from. However, in this case, unlike the case of political affiliations, we may choose to allow the possibility of a fake news detection engine that admits upfront that news from certain countries is likely to be judged with higher confidence than news from other countries.

The above examples suggest that extreme violations of *procedural fairness* along dimensions of political affiliation and regions should be considered as unacceptable. By way of *procedural fairness*, one would mean that the procedure for determining veracity should not be biased to favor some over others.

**Procedural Fairness in DFND** The desirability of procedural fairness places significant constraints on what kinds of DFND algorithms would be acceptable and which ones would be unacceptable. In other words, the notion of procedural fairness places constraints on how to go about building DFND methods. We will consider political affiliation as the dimension of consideration for fairness for this discussion; however, the ensuing discussion is equally applicable to any choice of dimension over which fairness is desired, such as gender, ethnicity, or geographical region. First, consider a purely data-driven DFND approach that is trained over historical labelled data. The notion of procedural fairness translates into fair representation of different political positions within each label in the training data. For example, if most fake news were from the right wing and most legitimate news were from the left wing, it would be easy for a learner to learn the (undesirable) mapping from political positions to a fake/real label. Even when representational parity is ensured, there is a possibility of algorithmic steps encoding some bias. Consider an example of a case where fake news from the left wing is more dispersed than fake news from the right wing. When regularizers are applied during the learning process for parsimonious model learning, the compact model may be inherently incapable of learning an accurate model to characterize the dispersed left wing fake news and thus would be able to deliver higher accuracies in detecting the more coherent right wing fake news. Thus, algorithmic steps including the usage of regularizers should be carefully scrutinized from the perspective of fairness. Second, for DFND methods that additionally incorporate external knowledge sources to inform decision-making, such sources should also be well distributed across political positions, with attention being paid to dispersion considerations as in the previous case. While the above checks do not yield a comprehensive fairness auditing method for DFND algorithms, a procedurally fair DFND method should necessarily align with the above principles.

**Impact Fairness in DFND** While we have seen that some forms of violations of procedural fairness would not be agreeable, it is interesting to consider what that entails for impact fairness. In general cases of using machine learning over person-level data, procedural and impact fairness are often in conflict. For example, if historical legacies of unfairness (such as racial unfairness) have resulted in significantly altered standing in terms of social, educational, and economic achievements across various categories (e.g., racial categories), a procedurally fair method would necessarily result in reflecting the biases. In other words, a race-agnostic and proce-

durally fair selection process could potentially result in much higher selection rates for whites than blacks if the former have a lower educational (and consequently, skill and achievement) profile, with race being correlated to the selection criterion of achievement level due to historical discrimination. Thus, in cases of handling person-level data, it is often argued, at least within progressive political circles, that impact fairness should take precedence over procedural fairness to counter historically entrenched discrimination manifesting as socioeconomic inequalities across dimensions over which fairness is desired. Streams of political philosophy such as Rawls [17], while stopping short of stating that impact fairness should trump procedural fairness, do prefer configurations where the inequality in impact is kept to as low levels as possible. With that background, the first consideration could be to ask whether impact fairness is in conflict with procedural fairness within the context of DFND. In other words, are there intrinsic or entrenched reasons as to why fake news is more abundant within a political position as opposed to another? In fact, while studies have generally been cautious about asserting a political correlation in fake news, there is increasing evidence that conservatives have historically played a much larger part in propagating fake news than liberals [10]. The study finds that extremely conservative people are almost twice as likely to spread fake news than extremely liberal people (these are self-reported labels, so need to be taken with a pinch of salt) on Twitter. Similar skewed distributions are potentially likely to be found when analyzing the kind of gender and ethnic stereotypes used in fake news authoring; for example, misogynistic fake news may be more prevalent than misandrist fake news. Evidence of such skewed distributions takes us back to the discussion on representation parity that we alluded to in the previous section. If indeed pro-conservative fake news is more prevalent than pro-liberal fake news, achieving representational parity requires us to sub-sample from available pro-conservative fake news, in order to construct a balanced training dataset. Even if such a balanced dataset is created, the method might still produce a significantly larger number of *fake* verdicts for conservative articles, since the skew along political alignment would exist in the unseen data over which these algorithms are used. The training data curation considerations are confounded in the presence of biases in user judgments (some bias mitigation strategies appear at [11]), especially when crowdworkers may have a different distribution of political positions as compared to that of fake news.

**Summary Remarks on DFND Fairness** We considered the two streams of machine learning fairness, those of individual fairness and group fairness. We observed that these are well motivated within the context of data about people but would still apply in some way in DFND. We considered concrete scenarios where DFND that is skewed on political and other positions may legitimately be considered as unacceptable. Building upon this, we observed that attention to procedural fairness along such dimensions would be necessary, and this places constraints and obligations on the design of DFND methods in myriad dimensions such as training data curation, algorithm design, and selection of external knowledge sources to use. Turning our attention to group fairness, we observed that such attention to

procedural fairness (that correlates to individual fairness) could lead to violations of group fairness which may be unavoidable against the backdrop of the observed skew of fake news distribution across dimensions such as political positions and gender. In summary, we may wish to place more emphasis on procedural fairness in DFND and pay significant attention to training data bias by ensuring well-designed and debiased data collection paradigms.

## 4 Democratic Values and Uptake of DFND

DFND, much like any classification task, is a computational labelling task. The ultimate goal is to pronounce a decision on each news article, which could be either a binary label of *fake* or *legitimate* or a veracity scoring on an ordinal scale. In typical data science scenarios, the predicted label is often used in a very straightforward way, that of associating the label with the data. In an automated job shortlisting scenario over received applications, the predicted label, which could be one of *shortlist*, *reject*, or *unsure*, could lead to concrete actions, such as sending a shortlisting or decline letter, or channelizing for more manual perusal of the job application. Guided by such commonly encountered settings, typical fake news detection software also uses a crisp verdict in the form of a predicted label. We saw this in the case of the NewsGuard application in Fig. 1 where the label is shown prominently at the top of the news article. We first take a look at the variety of ways in which fake news-related information (or in general, any kind of information veracity assessments) is presented to users, following which we will assess normative considerations within a liberal democratic framework and how they may favor some forms of presentations more than others.

**Current DFND-Based User Engagement** We will now consider a set of currently available software tools and algorithmic techniques with a focus on how they aid in tackling fake news and how they present veracity information to users, along with any forms of analysis they promote by way of the presentation modalities they use. One of the first efforts at developing a plugin for fake news detection was as early as 2016, leading to a tool called *BS Detector*.[7] The plugin has since been taken down and few traces of vivid details of its workings and resulting presentation modalities are available in the public domain. However, a screenshot available on the web seems to suggest that it places a veracity label on news articles and, at least in the case illustrated in Fig. 5, places an emphasis on the website from which the news is sourced. A similar software, *Stop the Bullshit*, illustrated in Fig. 6, squarely informs the user that the reasoning is based on the website, rather than the content that is attempted to be accessed. To illustrate the pervasiveness of website-level reasoning used in veracity result presentation, we may observe that Wikipedia

[7] https://www.bustle.com/articles/195638-the-bs-detector-chrome-extension-flags-the-sites-containing-a-load-of-political-bs.


**Fig. 5** BS Detector Plugin labelling a news piece as "This website is considered a questionable source" (picture courtesy: Bustle.com)

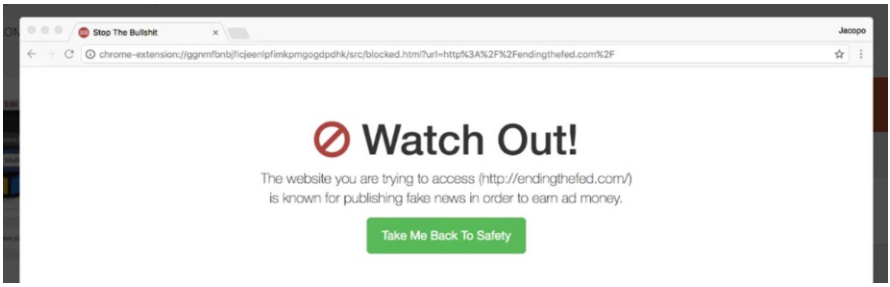


**Fig. 6** *Stop The Bullshit* software (picture courtesy: ProductHunt.com)

has a web page on *List of Fake News Websites*.[8] *Know News*, a veracity detection plugin by Media Monitoring Africa, which has been widely reported on the web as employing veracity detection on content, also predominantly uses website-level reasoning in presenting results (Fig. 7).

The website-level reasoning relentlessly expressed and promoted in veracity assessment presentation, we will see, may be critiqued from the perspective of alignment with democratic values. For now, it may be noted that such reasoning does not allow to recognize that the same website may host content of varying veracities. Despite the pervasiveness of website-level reasoning, there have been a few efforts that focus on the content and allow for veracity to be checked without using any

[8] https://en.wikipedia.org/wiki/List_of_fake_news_websites.

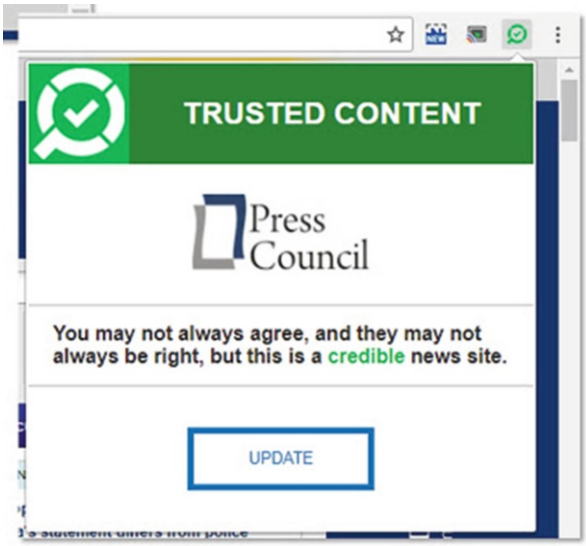


**Fig. 7** *Know News* plugin, a veracity assessment software, presenting veracity results basing the reasoning on the website (picture courtesy: Chrome Web Store)

kind of information about the source of the content. *ClaimBuster* [9], unlike the software that we saw so far, places an emphasis on automating the fact-checking of claims. While it calls itself an *end-of-end fact-checking system*, it limits itself to presenting pertinent information to a claim (mostly related claims) along with veracity information associated with such pertinent information. The user could then consider such information in arriving at a veracity judgment herself. Figure 8 illustrates the veracity results presented over a manually entered claim. By stopping short of providing a concrete and crisp veracity judgment, ClaimBuster places a higher cognitive burden on the user since the user has to digest and assimilate the related information presented in order to decide whether or not to trust the claim. However, in doing that, it also allows acknowledging the nuanced nature of veracity determination.

There have been several other "indirect" methods of tackling the fake news problem used by several stakeholders in the media ecosystem. These include enhancing *findability* of credible news sources. Facebook, the social media giant, explicitly admits using credibility/veracity information in the ranking of stories within news feeds. On one of its help pages,[9] it says: "Showing false stories lower in News Feed: if a fact-checker rates a story as false, it will appear lower in News Feed. This significantly reduces the number of people who see it." While this sounds quite reasonable, such credibility adjustments are notably done without

[9] https://www.facebook.com/help/1952307158131536.

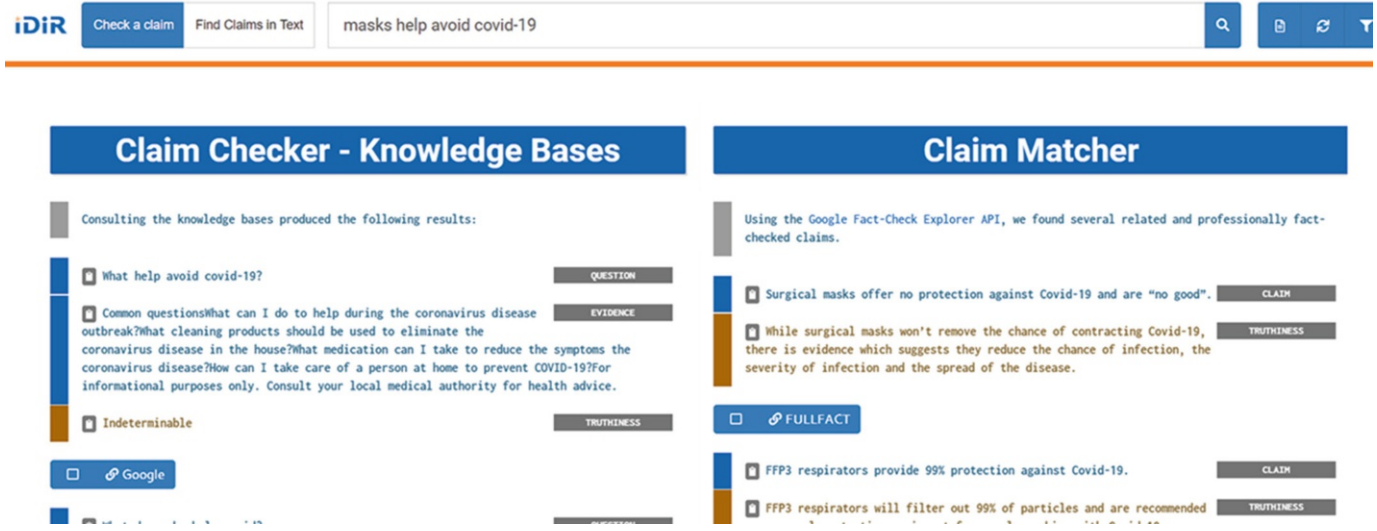


**Fig. 8** ClaimBuster in action; when a user enters a claim in free text, they find related claims and veracity information associated with them

user engagement. In particular, reducing the chances of seeing a lower credibility news story is sharply different from displaying a warning (like NewsGuard, BS Detector, and other examples seen before) since it marks a shift of agency (in the decision-making process of whether the article is to be read) from user to algorithm. The Facebook veracity/credibility judgments are on the basis of feedback from both users and what are called *third-party fact-checkers*. The same help page referenced above says: "Identifying false news: we identify news that may be false using signs like feedback from people on Facebook. Fact-checkers may also identify stories to review on their own." In particular, it says precious little on how an aggregate score on veracity is arrived at when there are conflicting signals from across users and fact-checkers and how a weighting is determined to balance user feedback and fact-checkers' judgments.

**How Should DFND Be Used?** Having looked at how DFND has been used, we now consider how DFND *should* be used. In this discussion, we draw heavily from the EU HLEG report on Disinformation [4] and look at high-level principles for DFND usage. Any way of ensuring that the results of DFND are put to use would result in some kind of barrier or constraint on the free consumption of all forms of information within society. Much like the institution of some binding norms could benefit everyone (e.g., traffic discipline helps everybody get to their destinations faster), it could be argued that enforcing binding norms that prevent creation and consumption of fake news could deepen democratic discourse. However, unlike traffic signal violations, there is an enormous amount of subjectivity in enforcing norms on media, to the extent that it would almost be impractical. The UN joint declaration on fake news [12], on the other hand, underlines the potential of fake news to mislead and interfere with the public's right to seek and receive, as well as impart information and ideas of all kinds. It also highlights the *positive obligation* of states to create an enabling environment for freedom of expression. This perspective, in contrast to the one illustrated earlier, puts forward a rights-based need for intervention. The EU report suggests that any disinformation interventions should be focused on two general objectives: (i) *increase long-term resilience* and (ii) *ensure that disinformation responses are up-to-date*. Let us now consider how these high-level and long-term objectives translate into the design of DFND usage within digital interfaces as well as elsewhere within society. The EU HLEG report also stresses on the importance of fake news responses to abide by five pillars, viz., *transparency*, *media and information literacy*, *empowering users and journalists*, *safeguarding the diversity and sustainability of the media ecosystem*, and *promoting continued research on the impact of disinformation*. These are summarized in Table 1.

**How Do Current DFND Methods Fare?** We now assess, qualitatively, as to how current DFND user uptake methods fare against the normative principles outlined earlier. The mainstream method of DFND adoption, that of assigning source-level (i.e., at the level of the website of the top-level domain) verdicts and presenting these to users, evidently does *not* conform to the *transparency* and *empowerment*

**Table 1** Recommendations from the EU HLEG report on disinformation [4]

| | |
|---|---|
| General objectives | Increase long-term resilience |
| | Keep disinformation responses up-to-date |
| Normative principles | Transparency |
| | Media and information literacy |
| | Empowering users and journalists |
| | Safeguard the diversity and sustainability of the media ecosystem |
| | Promote continued research |

criteria above and also may be seen as agnostic to safeguarding the diversity and sustainability of the media ecosystem. By explicitly indicating the verdict and allowing the user to disregard a *fake news* warning and continue reading, it may be said that there is some regard to media and information literacy, as well as user and journalist empowerment, within that model. The source/website-level verdict would need to be revised over time in order to satisfy the general objective of ensuring that disinformation responses are up-to-date; this would require that a website that has stopped sharing dubious content not be disadvantaged even after the change in character. In contrast to this analysis, reducing the findability of fake news by taking veracity into account in generating the ranking for the news feed, as used within Facebook, is quite weak in adherence to the normative objectives and principles laid out above. It may be argued that they neither satisfy the general objectives nor the five principles. In particular, such *under-the-cover*-type fake news exposure reduction methods, while sounding attractive in terms of offering a seamless integration into current systems, are quite poor when it comes to adherence to democratic and liberal values that have motivated most of the normative recommendations for fake news responses. The ClaimBuster approach, which starts with a claim and then presents related claims along with their veracity information, enabling a user to arrive at a judgment, can be seen as most amenable to the considerations seen above. However, as it was developed as a standalone tool which is not meant to intercept user-media interaction, its impact could be limited. The system, when augmented with a claim detection method, could well be packaged as a plugin which searches for relevant claims to the core claim on the web page attempted to be perused. That said, another drawback is that of the detailed nature of the presentation (i.e., related claims and their veracities), which makes it hard to be presented within a plugin format without significant detriment to user experience.

**Improving Adherence to Normative Principles in DFND Uptake** We now discuss how we could potentially improve the adherence to the principles from Table 1 in DFND uptake. Our attempt is not intended to outline a concrete and novel exemplary DFND approach, since the development of such a framework would naturally take several years of research effort. However, we will attempt to outline recommendations based on currently available technologies in order to translate the high-level principles into a language that is better understood by technologists in

machine learning and data science. In particular, the aspects of *transparency* and *user empowerment* are quite interesting to analyze in terms of how they could be realized computationally. We outline some high-level components of a roadmap toward enhancing adherence to the principles outlined earlier:

- **Relaxing (Implicit) Obligations of Showing a Crisp Decision:** Virtually all veracity-oriented software and tools do pronounce a crisp decision on what is evaluated. This is likely almost construed as entailing from the task undertaken by the tool. However, we may argue that there is no need to show a crisp decision as such. Showing critical information that would empower the user to arrive at a decision for herself could be considered as *enough*. We realize that such a tool may lose out on *user appeal*, and the ability to show a crisp and clear decision is often part of the *hard sell* marketing that DFND tools may use; thus, market forces are likely not conducive toward relaxing the paradigm of showing a crisp decision. Relaxing the paradigm of crisp decision-making would require the interested user to engage better with the information being presented and thus could empower users. Furthermore, crisp decisions that are mostly informed by source-level (i.e., website-level) cues could be also seen as a soft censorship and are thus not well aligned with the goal of safeguarding the diversity and sustainability of the media ecosystem.
- **Confidence Scores with Decisions:** The attractive feature of offering a crisp decision is often used without an understanding of whether such decision-making is valid; in other words, we seldom ask the question *have we designed the algorithm to tell us when it does not know enough* to make a decision. Often times, the decision-making is made in a comparative manner, based on which is the best choice among available decisions. This obscures information on whether the algorithm is indeed confident about the decision it is making or whether it is the best effort choice made under considerable ambiguity. DFND methods, given the importance of the domain of operation, need to have both a mechanism of reporting the *error bar* in some intelligible manner and a probability score (perhaps expressed as a percentage). Such confidence scores help deepen the adherence to the normative principles in Table 1.
- **Explaining Decisions:** There has been much recent interest in explainability in AI and machine learning (e.g., [1]). Enhancing explainability in fake news detection is a direct way of enhancing transparency, as well as media and information literacy. Neural network-based models have often been criticized for lack of understandability of the decision-making process, and enforcing a condition of explainability within the learning process may cause them to operate at lower levels of accuracy. Thus, explainability could stand in the way of achieving the best possible accuracies in decision-making, creating an interesting trade-off between accuracy and democratic values. Another way of tackling the problem is to do post hoc explanations, whereby the explanation is made after the decision-making. Cases when a post hoc explanation cannot be derived could be indicated explicitly, so the user may treat that as an indicator of having to take the decision with a pinch of salt.

- **Showing Pertinent Credible Information:** For systems that make use of sources of credible information (e.g., PubMed articles in the health domain, ontologies for science, and so on), a straightforward way to use them would be to display *pertinent credible information that would enable manual verification* directly to the user. This would enhance user engagement while also implicitly training users to exercise own judgment and analysis, something that could be deemed critical for long-term resilience. Showing information from known credible source is related to, but different from, the ClaimBuster approach of showing related claims in the dataset along with their labels.
- **Showing Pertinent Non-Credible Information Marked Clearly:** A complementary approach to showing credible information would be to show fallacious/fake information marked so, as long as it is related enough to the article upon which a decision is to be made. This paradigm, one may argue, might nudge the user to engage in fact-checking or verification, by showing that there is quite similar content that is known to be fake. However, this paradigm needs to be used with abundant caution due to several reasons. First, showing non-credible information to users enhances user familiarity with such content, and this heightened familiarity along with the psychological bias called *illusion of truth effect*[10] could eventually lead to an enhanced belief in such fake news. Second, showing such non-credible information in an easily accessible manner might risk the DFND method being perceived as an easy-access channel for fake news. One way to mitigate such risks could be to mark such non-credible information very clearly, as illustrated in Fig. 9.
- **Encouraging User Engagement and Deeper Analysis:** In addition to the above, DFND methods may explore interactive tools in order to enhance user engagement and empowerment. For example, the suggestions above of showing credible information, explanations, and confidence scores could all be operationalized using a mouse hover paradigm. For example, hovering the mouse over a particular sentence could bring up a tooltip with information localized to the sentence. DFND methods could also employ force-directed graph-based interfaces[11] which show the interrelationships between segments of the article in an interactive manner to aid user-specific explorations to enable deeper understanding of the veracity judgments presented.
- **Ability to Provide Feedback and Other Information:** DFND methods, much like any machine learning method, are not designed to achieve perfect decision making, and could make wrong decisions. Thus, it would be in the interest of the DFND method to continuously improve the decision-making processes through crowdsourcing feedback from own users. In addition, the facility to provide feedback and other kinds of information (e.g., credible information pertinent to the article in question) will hopefully enhance user confidence in the system and promote media literacy and user engagement.

[10] https://en.wikipedia.org/wiki/Illusory_truth_effect.

[11] https://en.wikipedia.org/wiki/Force-directed_graph_drawing.


**Fig. 9** Abundant caution is necessary while displaying fake news to users, even if it may be to debunk it. The picture shows how AltNews, an India-based fake news detection engine, uses several ways of indicating that the information is fake while displaying it to the user

The above recommendations are not meant to be comprehensive, but we hope this will enhance and sharpen the debate on how AI should go about tackling fake news while acknowledging, preserving, and deepening the democratic values within society.

## 5 Conclusions

In this chapter, we analyzed and discussed several ethical and normative considerations that are relevant to the context of data-driven and AI-based automation of fake news detection. We began by outlining the increasing pervasiveness of fake news detection and its societal and political importance, motivating the need for increased attention to the non-technological aspects of data-driven fake news detection (DFND). We also analyzed the main pillars of ethical considerations. First, we considered the historical context of machine learning, and argued that legacy considerations of optimizing for automation-era metrics are squarely inappropriate to drive its evolution within domains such as digital media within which DFND

is situated. Second, we considered why the very nature of data-driven learning could be critiqued for usage in domains within volatile dynamics within which automation could be an active player. Third, we described that the domain of fake news detection has some unique features which make it unlike other analytics tasks with a similar structure such as spam detection and product recommendations. Following this, we delved deeper into fairness considerations in DFND. We analyzed DFND from the perspectives of the two streams of fairness concepts used within ML, viz., individual and group fairness. We argued that individual or procedural fairness may be considered as being more important for DFND and contrasted it within analytics involving person-level data where group fairness may legitimately be considered more critical. We then turned our attention to analyzing DFND uptake modalities and how they fare against recent recommendations on normative principles that DFND should align with. We observed that under-the-cover and seamless integration of DFND results, while sounding attractive, would fare significantly worse on such normative principles, as opposed to more explicit ways of delivering the results to the user. Based on such analyses, we outlined several technologically grounded recommendations that could inform the design and development of DFND methods that would be well aligned toward preservation and deepening of democratic values within society.

## References

1. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fus. **58**, 82–115 (2020)
2. Bernardin, H.J., Beatty, R.W., Jensen Jr., W.: The new uniform guidelines on employee selection procedures in the context of university personnel decisions. Person. Psychol. **33**(2), 301–316 (1980)
3. Binns, R.: On the apparent conflict between individual and group fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 514–524 (2020)
4. Buning, M.D.C., et al.: A multidimensional approach to disinformation. EU Expert Group Reports (2018)
5. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226 (2012)
6. Freyenhagen, F.: Taking reasonable pluralism seriously: an internal critique of political liberalism. Polit. Philos. Econ. **10**(3), 323–342 (2011)
7. Gangireddy, S.C.R., Long, C., Chakraborty, T.: Unsupervised fake news detection: a graph-based approach. In: Proceedings of the 31st ACM Conference on Hypertext and Social Media, pp. 75–83 (2020)
8. Hashemi, M., Hall, M.: Criminal tendency detection from facial images and the gender bias effect. J. Big Data **7**(1), 1–16 (2020)
9. Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A.K., et al.: Claimbuster: the first-ever end-to-end fact-checking system. Proc. VLDB Endowm. **10**(12), 1945–1948 (2017)

10. Hopp, T., Ferrucci, P., Vargo, C.J.: Why do people share ideologically extreme, false, and misleading content on social media? A self-report and trace data-based analysis of countermedia content dissemination on Facebook and twitter. In: Human Communication Research (2020)
11. Hube, C., Fetahu, B., Gadiraju, U.: Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2019)
12. Huff, M.: Joint declaration on freedom of expression and "fake news," disinformation, and propaganda. Secrecy Soc. **1**(2), 7 (2018)
13. Hysolli, E.: The story of the human body: Evolution, health, and disease. Yale J. Biol. Med. **87**(2), 223 (2014)
14. Lloyd, E., Wilson, D.S., Sober, E.: Evolutionary mismatch and what to do about it: a basic tutorial. Evolut. Appl. 2–4 (2011)
15. Narayanan, A.: *How to Recognize AI Snake Oil*. Arthur Miller Lecture on Science and Ethics. MIT, Cambridge (2019)
16. O'neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Broadway Books, New York (2016)
17. Rawls, J.: A Theory of Justice. Harvard University Press, Cambridge (1971)
18. Roets, A., et al.: 'Fake News': incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. Intelligence **65**, 107–110 (2017)
19. Smart, J.J.C., Williams, B.: Utilitarianism: For and Against. Cambridge University Press, Cambridge (1973)
20. Thurman, N.: Personalization of news. Int. Encycl. Journal. Stud. **2019**, 1–6 (2019)
21. Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., Cave, S.: Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research. Nuffield Foundation, London (2019)
22. Yin, X., Han, J., Philip, S.Y.: Truth discovery with multiple conflicting information providers on the web. IEEE Trans. Knowl. Data Eng. **20**(6), 796–808 (2008)