AI & SOCIETY
https://doi.org/10.1007/s00146-025-02324-8

RESEARCH

The Limits of Machine Learning Models of Misinformation

Adrian K. Yee[1]

Received: 26 February 2025 / Accepted: 14 March 2025

Abstract
Judgments of misinformation are made relative to the informational preferences of the communities making them. However, informational standards change over time, inducing distribution shifts that threaten the adequacy of machine learning models of misinformation. After articulating five kinds of distribution shifts, three solutions for enhancing success are discussed: larger static training sets, social engineering, and dynamic sampling. I argue that given the idiosyncratic ontology of misinformation, the first option is inadequate, the second is unethical, and thus the third is superior. However, I conclude that the prospects for machine learning models of misinformation are far weaker than most have presupposed, given that both epistemic and non-epistemic values are difficult to operationalize dynamically in machine code, rendering them surprisingly at most a species of recommender systems rather than literal truth detectors.

Keywords Misinformation · machine learning · philosophy of social science · philosophy of science · social epistemology

1 Introduction

Misinformation has recently been a topic of discussion in the intersection of philosophy of science and artificial intelligence via 'machine learning models of misinformation' (MMMs): the usage of computers to automate the process of identifying misinformation in text and images (Yee 2023a, b; Harris 2024). MMMs are of central interest to citizens and governments, with some advocating their usage for regulating societies' information ecosystems, such as in Malta (Cassar 2023) and Saudi Arabia (Altiabi 2022). However, the overall adequacy of MMMs remains highly contested, despite scoring highly on standard machine learning metrics of accuracy, precision, and recall (Cf. Khan et al. 2021). Given how widespread misinformation is alleged to be, and given the high stakes that governments, citizens, and private industry have in ensuring that information flowing through the internet and other sources is as high quality as possible, it is important to reflect on the increasing usage of MMMs and analyze their methodological foundations.

This paper focuses on a neglected methodological problem in MMM research. Judgments of misinformation are always made relative to a background epistemic community whose informational norms are in sufficient equilibrium: an act of communication is judged as misinformation whenever violations of epistemic conventions surrounding usage of that kind of information are perceived to have been made, typically concerning the truth or misleadingness of a piece of information. However, empirical and historical studies have shown that sufficient equilibrium is often merely transient concerning prevailing informational norms, rendering these systems the result of stochastic processes whose non-stationarity is particularly difficult to model. This suggests that when one tries to automate manual judgments of misinformation in the form of MMMs, any ostensible successes of test data relative to training data become underwhelming when there are strong empirical reasons to be skeptical of the success of real-world applications due to what been called the MMM problem of 'temporal generalizability' (Stepanova and Ross 2013). While the problem of distribution shifts in machine learning more generally has been studied in great detail (Gama et al. 2014), the context of MMMs remains comparatively under studied with unclear causes and solutions for such shifts.

As I will argue in this paper, social and computer scientists ought to reconsider what the appropriate goals and methodological limits of MMM construction are, namely that they cannot be justified as literal falsity detectors as most social scientists presently conceive of them, but instead

✉ Adrian K. Yee
adrianyee@ln.edu.hk

[1] Lingnan University, Hong Kong Catastrophic Risk Centre, Tuen Mun, China

Published online: 09 May 2025

are intrinsically sensitive to the social and epistemic context in which they are deployed (Søe 2018). Failure to heed under analyzed features of judgments of misinformation present practical difficulties for automated detection of misinformation in internet ecosystems cross-culturally. This is especially so given that international disputes about what counts as misinformation concern metaphysical, scientific, and political matters of significant disagreement concerning the role of epistemic and non-epistemic values in judgments of misinformation.

Section 1 begins with examples illustrating the different ways in which judgments of misinformation are typically given, setting the scene for understanding MMMs. Section 2 discusses current issues in MMMs and Sect. 3 introduces five kinds of distribution shifts that can confound empirical adequacy. Section 4 then discusses three ways to improve the empirical adequacy of MMMs and illustrates the confounding role of epistemic and non-epistemic values in inducing significant distribution shifts regarding judgments of misinformation over time.

2 Understanding judgments of misinformation

Before articulating the methodology of contemporary MMMs, it is important to first understand how judgments of misinformation typically have functioned in the contexts of history and daily life. Societies differ considerably in how they impose standards and set expectations for what is considered good informational discourse. These can be broadly categorized into four features: false information, misleading information, disputes about epistemic values, and disputes about non-epistemic values.

Misinformation is often politicized, as when news organizations or governments spread false information to persuade target audiences. To use a modern historical example, the Japanese government taught deliberate distortions of Buddhist and Shintoist doctrines in the early half of the 20th-century to convince Japanese citizens that they were the 'master race' and that their imperialist activities were justified, insofar as massacres would 'cleanse the region of non-Japanese people' (Rigg 2023). In this case, the Japanese government systematically taught false propositions to convince its citizens and members of the Imperial Japanese Army, that it was permissible to commit innumerable war crimes leading to the deaths of millions of people throughout Asia.

In other cases, the context surrounding an utterance of a proposition renders it a form of misleading misinformation, rather than the propositional content itself. For example, a friend discourages another from vacationing in Australia on the grounds that Australia has dangerous and deadly creatures. This is accurate but highly misleading insofar as it is improbable such creatures would harm most tourists vacationing in major cities in Australia. Here, it is the context of utterance (i.e., that one is advising a friend on vacation destinations), rather than the propositional content itself, that is playing more of a role in the assessment of whether the utterance is misinformation. To use another example, publicly expressing strong dissent against the Singaporean government can be considered illegal misinformation, insofar as it is considered seditious and a threat to social cohesion in that country (Republic of Singapore 2019). But identical such remarks uttered in societies such as the United States are not only legal but not categorically considered misinformation. Hence, what counts as misinformation is relative to the social goals of a community's members, and identical propositional content does not ensure identical status as misinformation. This is especially salient when one considers the fact that MMMs are developed in societies whose citizens often have diametrically opposing metaphysical and ethical views (e.g., Saudi Arabia versus Canada), introducing substantial methodological issues when constructing fake news and misinformation classifiers cross-culturally (Touahri and Mazroui 2024).

To use an example of how epistemic values play a role, consider the claim that 'COVID-19 vaccines cause harmful negative side effects and are therefore dangerous if used'. While there are many ways to understand this claim, one may try to argue that the probability (e.g., 3%) of a given vaccine producing negative side effects is considered too high for someone's personal risk tolerance. But others with differing risk tolerances would consider this claim misinformation, insofar as they may have a much higher risk tolerance and protest that 3% is sufficiently low enough that it is misleading to claim that such vaccines are dangerous. Here, an individual's risk tolerance plays a central role in adjudicating what counts as misinformation, where risk tolerance involves both a probability (an epistemic value) and a set of possible negative outcomes (a non-epistemic value).

To use an example of how non-epistemic values play a role, social scientists have recently argued that many misinformation researchers confuse the prevalence of 'internet memes' for widespread dissemination of misinformation that harms people, in which sometimes hateful, false, or misleading imagery is actually shared with comedic rather than malicious intent (Altay et al. 2023). Here, value-ladenness is intrinsic to the judgment of whether an item of information is misinformation insofar as the boundaries of what are considered humor versus sincere political commentary are disambiguated only by the risk-weighted judgments of stakeholders participating in or observing these discourses. Sometimes content perceived as jokes by one community is perceived by another sub-community as oppressive semiotic systems reinforcing misinformation through stereotypes (e.g., racist online humor). That is, those who consider

certain internet memes offensive may judge them as such given that they believe that dissemination of this information may lead others to adopt racist attitudes, rather than perceive them as possibly poor taste humor. A key factor here is whether information is being presented from sources that cohere with the epistemic expectations of those using the information source and in a manner that is political correct, where what counts as politically correct is a matter of non-epistemic values of appropriateness, offensiveness, and tactfulness.

Despite these examples, there remains no agreed upon definition or theory of how to understand the varieties of misinformation. Nonetheless, mainstream European law and government policy documents have typically defined *misinformation* as false information disseminated without intended harm, *disinformation* as misinformation with the intent to harm, and *malinformation* as true information intended to harm (Fathaigh et al. 2021, 4). I call this family of views the 'alethic theory of misinformation' insofar as the truth-value of a proposition is the core constitutive feature of what makes information misinformation. Nearly all scholars of misinformation advocate some form of alethic definition of misinformation including philosophers (Floridi 2011; de Ridder 2021), MMM theorists (Rafiqul et al. 2020), and most surveyed social scientists (Altay et al. 2023), with but a few advocating alternative non-alethic frameworks (Yee 2023a, b; Swire-Thompson and Lazer 2020).

However, alethic theories have recently been shown to be methodologically problematic: misinformation cannot have as either a necessary or sufficient criterion for its rational application as a term that information be 'false' or 'misleading'. It cannot be defined merely as false information, as this would return the unreasonable verdict that all previous false scientific theories (e.g., Newtonian mechanics or bloodletting) are misinformation, even if such theories adhered to the highest epistemic standards of their time. Similarly, this alethic account would have the verdict that nearly all contemporary sciences using regression methods, given their non-trivial error bounds, are a form of misinformation, which is absurd (Yee 2023a). It also cannot be defined merely as misleading information because this would entail that magicians, any pedestrians wearing makeup, and professional comedians all produce misinformation, even though all stakeholders (e.g., target audience members) not only consent but desire to be misled in ways that are esthetically pleasing, entertaining, or cause no harm. Hence, some forms of intentional misleading information are even of positive social utility. Furthermore, atheists and religious people fundamentally disagree on whether religious texts constitute misinformation so defined (e.g., what counts as misinformation in the United States is not the same as in Saudi Arabia), and cultures of radically distinct epistemic backgrounds will disagree as to what they consider true or false propositions, rendering alethic theories of misinformation often intractable to use for policy purposes (Yee 2023b). On the grounds of the existence of widespread religious disagreement alone, it is methodologically futile to employ an alethic account when training MMMs given how many religions do not agree with one another on the basic structure of our universe and human ethics, as this would entail widespread judgments of misinformation relative to one religious communal standard vis-a-vis another.

Instead, a more reasonable methodology to employ is to consider that the violation of *additional* informational preferences is required for a judgment of misinformation to be misinformation, rather than merely the extent at which the information is false or misleading. For example, a high school physics teacher teaching classical mechanics is not engaged in disinformation, because this curriculum satisfies the relevant informational preferences of stakeholders: teenagers are learning instrumentally and explanatorily powerful theories of physics even if they contain false idealizations (e.g., mass is invariant with respect to acceleration). While the falsity or misleadingness of information is undoubtedly some component of what makes information misinformation, it is far from clear that it is the most fundamental component as compared with whether an item of information is relevant, presented in a manner that meets expectations given the source and format, satisfies broader social goals surrounding appropriate usage of that information, or meets the general epistemic standards of the community (e.g., provides explanatory power, is parsimonious, consonant with background non-epistemic values, etc.). For instance, empirical data from Osman et al. (2022), surveying several thousand respondents from Russia, Turkey, United Kingdom, and United States, suggested that 69% consider 'misinformation' as 'information that is intentionally designed to mislead' (rather than unintentionally), 49% that it is information that 'exaggerated conclusions from facts', 48% that it 'didn't provide a complete picture', and 43% that it was 'presented as fact rather than opinion or rumour'. In the words of some health researchers studying medical misinformation, misinformation is better understood as "information that is contrary to the epistemic consensus of the scientific community regarding a phenomenon...what is considered true and false is constantly changing as new evidence comes to light and as techniques and methods are advanced" (Swire-Thompson and Lazer 2020, 434). These latter theories more adequately respond to the context sensitivity of judgments of misinformation than do mainstream alethic theories.

Despite the aforementioned methodological challenges, the majority of MMM researchers nonetheless continue to adopt alethic theories of misinformation and employ a working definition of misinformation as 'false or misleading information' (Rafiqul et al. 2020, 81; Caled and Silva 2022, 126–127). Against this social scientific mainstream,

I will argue that this methodology of employing an alethic definition is even harder to justify in the actual practice of MMM development.

3 Machine learning models of misinformation

There are broadly two categories of MMMs: those which analyze propositionally structured misinformation and those which analyze non-propositionally structured misinformation. As an example of the former, unsupervised learning techniques have been used to associate emotionally valent words in sentences judged to be misinformation (e.g., containing false or misleading propositions) using convolutional neural network architectures (Lumvembe et al. 2023). As an example of the latter, supervised learning algorithms have discerned certain features of 'fake photographs' (i.e., photographs which have no causal connection between what they depict and reality) that perform well on test photographs, and often perform at least as well as human beings' abilities to discern fakes (Groh et al. 2022). Several classifiers have been constructed identifying fake images, fake news, false propositions, and other forms of deceptive media.

The general procedure for designing MMMs can be summarized as follows (Yee 2023b, 7). There is typically a division between those advocating an MMM, those designing the MMM, what data the MMM is trained on, those labeling data in supervised learning contexts, who is deploying the MMM, and the relevant stakeholders that the MMM's outputs are intended to apply to. For example, a group of medical researchers attempting to determine the frequency of homeopathic remedies for COVID-19 (e.g., a form of medical misinformation) on Facebook posts are both advocates and stakeholders in the output of an MMM. Others might then design the MMM, gather data drawn from Facebook's 'application programming interface' for training the algorithm (e.g., hyperparameter and weight optimization), and clean, organize, and label data, with eventual deployment on test data distinct from training data. The informational standard set regarding what counts as misinformation is decided by the medical researchers themselves and thus determines any operationalizations of ground truth determined by such standards. Typically, this standard is manifested by instructing digital laborers (e.g., hired from Amazon Mechanical Turk) to adhere to the first-order judgments of those the medical researchers consider epistemic elites, such as journalist Craig Silverman's widely circulated BuzzFeed list of fake news websites, when labeling data that will eventually be used to set MMM parameters (Cf. Guess et al. 2019).

Determining the empirical adequacy of MMMs remains challenging for several reasons, such as determining the extent at which they can successfully predict novel or future misinformation and whether the set of ground truths is sufficiently stable over time such that a classifier makes judgments of misinformation similar to what a community's ideal informational agent would similarly judge. After all, an MMM is supposed to automate what would otherwise be the manual judgments of informational agents operating in accordance with communal epistemic norms. This is the sole standard of empirical adequacy at play given that misinformation is an intrinsically value-laden phenomenon that is the result of social construction, rather than inhering in nature itself. Hence, there is a legitimate plurality of reasonable first-order judgments as to what items of information are considered misinformation. What is considered misinformation in one country or political context is not in another, and it is unclear how to adjudicate such claims' adequacy on epistemically robust grounds. This has led some to argue that adjacent concepts to misinformation, such as 'fake news', often have too vague or unstable referents to be of practical usage by researchers or policymakers (Habgood-Coote 2019, 1039–1040). Indeed, recent empirical evidence suggests significant disagreement as to the misinformation judgments of prominent fact-checking organizations, such as Snopes and PolitiFact, with as much as 228 (30.4%) out of over 749 articles sampled on identical topics receiving diverging ratings of informational quality (Lee et al. 2023, 2).

Indeed, previously discussed issues with alethic theories of misinformation are particularly salient in the context of constructing MMMs. It is unreasonable for an MMM to be assigned the task of automating the separation of true information from false information, as this is tantamount to attempting to construct an impossible magic oracle deciding the truth-value of all propositions. Hence, MMMs cannot have as their literal goal the detection of either false or misleading propositions or news, given that we are often not in a position to know the truth (e.g., forecasting future political events). At most, MMMs can automate the classification of what an epistemic community *judges* to be false or misleading information: "This requires the researcher to either pass their own judgment on the veracity of news or to rely on outside journalistic organizations for labeling" (Horne et al. 2020, 2). But even on this account, Yee (2023b) has argued that the construct legitimacy and construct validity of supervised MMMs has often been contentious because labeling procedures are typically conducted by non-representative stakeholders in the adjudication of information quality: a construct C is *construct legitimate* if and only if C has properties that are justified by a background scientific theory, and a machine learning classifier M is *construct valid* with respect to a construct C if and only if M adequately tracks C in the sense of capturing C's content appropriately when operationalized in machine code. Problems for both construct legitimacy and construct validity arise in MMM model construction given the role of epistemic elites in

labeling data (in supervised learning contexts) and the provision of curated unstructured data (in unsupervised learning contexts). MMMs trained in this fashion can, therefore, lead to a problematic epistocracy where information ecosystems are curated by epistemic elites with their own biases that can be reinforced through an MMM's outputs, without considering the non-trivial role that citizens should have in adjudicating aspects of information quality (Yee 2023a; Horne et al. 2023, 15).

4 The problem of distribution shifts

While I have highlighted several concerns with the methodology of MMMs, what more severely compromises the empirical adequacy of MMMs is the comparatively neglected problem of distribution shifts.

> **Distribution Shifts**: Models of misinformation are constructed from samples drawn from stochastic processes[1] that fail to be weakly stationary, given that what is considered misinformation changes radically over time and context.

A variation of this problem has been raised previously in the MMM literature: "[G]round truth will guide the algorithm to distinguish between true and false content, making training possible. However...an objective ground truth might not exist, or might change over time" (Horne et al. 2023, 6). Indeed, the authors add that there cannot be a proper science of MMMs without stronger statistical assumptions holding in the real-world deployment of the machine learning model: "[W]e can only make accurate predictions when the new data points given to the system are independent and identically distributed [IID]...While this assumption is fundamental to ML and predictive analytics, it is often ignored in fake news detection research" (Horne et al. 2023, 6). However, the claim about IID is too strong because one can still make reasonably actionable predictions about the future states of stochastic processes that are weakly stationary, insofar as effective policies are possible while knowing only the mean and variance of the distribution of judgments of misinformation (i.e., without requiring stability of the skewness, kurtosis, and higher moments of the stochastic process's moment generating function). However, as I will argue, even this standard of weak stationarity is unlikely to hold over time in many real-world social contexts.

To illustrate the problem, consider first as a foil to real-life the following ideal societal arrangement in which an MMM would be easier to implement. Such an informational society would be one in which there was no disagreement about informational norms: there is no disagreement about which items of information are judged as misinformation, and thus the mean and variance of these judgments would be constant over time. Thus, one could simply sample any arbitrarily small subset of this population to estimate population-level judgments of misinformation to train an MMM to make similar judgments as the manual judgments of community members. Furthermore, suppose that this community does not change their informational preferences and thus retain the same informational standards over time. In this sense, any MMM designed this way would reflect this society perfectly and thus would have perfect empirical adequacy on test data, given the training data is trivially representative of the entire population by virtue of this community's homogeneously distributed informational preferences.

However, in reality, this is never the case and we have significant divergence from this idealized society synchronically and especially diachronically. Before providing more specific examples of distribution shifts in MMS, it is important to emphasize the severity of what is at stake in constructing MMMs: the computer automation of any human task can lead to problems resulting from the speed and scalability of computing systems. MMMs will undoubtedly continue to be used in one format or another to regulate internet ecosystems for hate speech, censorship, moderation, and propaganda. In doing so, MMMs introduce three kinds of risks that accrue to the high velocity and degree at which concepts are implemented in machine code: (i) that a failure of construct legitimacy risks quickly and severely affecting information ecosystems with harms that would not scale if conducted via merely human manual procedures; (ii) that a failure of construct validity can lead to MMMs creating information ecosystems that diverge significantly from the intended constructs that we trained our MMMs with, thereby alienating ourselves from such systems; (iii) that any errors in the methodology of MMMs will percolate to other machine learning systems that they may be operative together with (e.g., large language models (LLMs) querying information or news aggregators filtering our 'fake news'). Relatedly, it has been recently argued by Yee (2025) that there are catastrophic risks whenever there is a failure of construct legitimacy and construct validity in machine learning models of counterterrorism, given that automating such procedures involves deadly consequences for those targeted by foreign military intelligence agencies assassinating alleged terrorist targets. The stakes are similarly high for misinformation detection given that, for example, people's lives are at stake when it concerns the identification and potential censorship of what counts as

[1] A stochastic process, understood as a random variable distributed over time, is *weakly stationary* if it is time-invariant with respect to its mean and covariance, and 'strongly stationary' if it is time-invariant with respect to any finite lag in all of its higher moments (e.g., skewness, kurtosis, etc.) (Hamilton 1994, 43–46).

medical misinformation. Misinformation can lead to fraud, financial losses in the stock market due to scams, physical and mental health issues, and more. Hence, it is necessary to understand what processes can disrupt the flow of MMM model construction and lead to less predictively accurate or explanatorily adequate models.

4.1 Five kinds of distribution shifts in MMMs

I now provide detailed empirical evidence for distribution shifts in judgments of misinformation and distinguish five kinds of MMM distribution shifts using a simplified example of a binary misinformation classifier over some set of input covariates. Letting **x** be a vector of covariates (e.g., a sentence in natural language or the set of pixels in a deep fake image), Y a random variable returning '1' for misinformation detected and '0' otherwise, $P_{train}(\mathbf{x})$ be the probability of covariates on training data and $P_{test}(\mathbf{x})$ on test data, '$\triangle$' some measure of divergence between two probability distributions, and '$\approx$' meaning 'approximately equals', we have:

Covariate Shift A (**CSA**): Covariates change such that $P_{train}(\mathbf{x}) \triangle P_{test}(\mathbf{x})$ but $P_{train}(Y|\mathbf{x}) \approx P_{test}(Y|\mathbf{x})$.
Covariate Shift B (**CSB**): Concepts change such that $P_{train}(\mathbf{x}|Y) \triangle P_{test}(\mathbf{x}|Y)$ but $P_{train}(Y) \approx P_{test}(Y)$.
Label Shift (**LS**): Labels change such that $P_{train}(Y) \triangle P_{test}(Y)$ but $P_{train}(\mathbf{x}|Y) \approx P_{test}(\mathbf{x}|Y)$.
Concept Shift(**CS**): Concepts change such that $P_{train}(Y|\mathbf{x}) \triangle P_{test}(Y|\mathbf{x})$ but $P_{train}(x) \approx P_{test}(x)$.
Radical Shift (**RS**): Covariates, labels, and concepts all change such that $P_{train}(\mathbf{x}) \triangle P_{test}(\mathbf{x})$, $P_{train}(\mathbf{x}|Y) \triangle P_{test}(\mathbf{x}|Y)$, and $P_{train}(Y|\mathbf{x}) \triangle P_{test}(Y|\mathbf{x})$.

I employ a relation of approximation '$\approx$' instead of equality '=' between terms given that in practice, probabilities are never equal but are close enough to one another relative to the practical purposes of social scientists and policymakers, as determined by a variety of epistemic and non-epistemic values featuring in model selection (Yee 2023b). In addition, there are at least two ways one can measure divergences '$\triangle$' between probability distributions so as to estimate the degree of distribution shift. The first is to simply look at classifier performance, such as accuracy, precision, F_1-score, etc. where, for instance, poorer performance during cross-validation can sometimes suggest a distribution shift of some kind is occurring. A second method is to measure the Kullback–Leibler (KL) divergence between the test and training distributions, which is a common measure of the dissimilarity between two distributions.[2] However, since the KL divergence is only well-defined if the input space is the same between both distributions, it cannot be used in many cases in which RS applies.

In what follows, I provide five salient empirical examples in the MMM literature illustrating each of CSA, CSB, LS, CS, and RS respectively.

4.2 Covariate shift A

Covariate shift A (CSA) occurs whenever the frequency of specific covariates changes between training and test contexts, but the conditional probability of these respective sets of covariates being labeled as misinformation is approximately the same in both contexts. For instance, Horne et al. (2020b) train classifiers on data from what were considered reliable news sources in the US and the UK, as well as unreliable news sources regardless of location. Calibrating their model with respect to the 'factuality scores' of Media Bias/Fact Check as ground truth, they randomly sample 100 articles from a larger sample of each country's set of news sources, and employ a variety of standard algorithms such as random forest and support vector machines. Surprisingly, the classifier methods struggled to perform well on test data, with the authors reporting that they can "partially attribute the trouble in classifying unseen, unreliable sources to the wide range in writing styles across these sources" (3). Furthermore, combining both the UK and US training data did not help to increase empirical adequacy, as measured by accuracy. They conclude that classifiers detecting misinformation trained on datasets in one country (e.g., US news feeds) do somewhat poorly when applied to other country's news feeds, even if the data is in the same language (e.g., UK news feeds). Hence, despite ostensible structural similarities in two populations, idiosyncrasies between two dialects of a language can seriously confound classifiers' predictive powers. Therefore, different kinds of covariates had a role to play in the classifier labeling them misinformation, but that relative to each distinct set of covariates, their conditional probability of being labeled misinformation was approximately the same (i.e., both the UK and US English news datasets had similar distributions in terms of their 'factuality scores').

4.3 Covariate shift B

As an example of covariate shift B (CSB), consider the recently proposed Time-Dependent Hawkes process model articulating the dynamics of fake news spread on Twitter (Murayama et al. 2021). This supervised learning classifier

[2] This is defined such that, for two discrete probability distributions p and y defined on the same space $\mathcal{X}$, $KL(p|y) := \sum_{x \in \mathcal{X}} p(x) log(\frac{p(x)}{y(x)})$. This generalizes to the continuous case, where we swap the summation for an integral. However, this is not a metric in the technical sense because it is an asymmetric function and violates the triangle-inequality. Nonetheless, it suffices as a measure of divergence for the purposes of the IS score.

was trained on labeled tweets fact checked by the websites Politifact.com and Snopes.com between March and May in 2019. A wave equation is constructed of probabilities: an initial peak in the frequency denotes misinformation first spreading followed later by a smaller peak in which users attempt to correct that piece of misinformation by re-sharing it while simultaneously adding corrective information. This wave equation is claimed to capture highly generalized properties of "the future evolution of the spreading of fake news on Twitter" (13).

However, this model makes the strong assumption that the relevant covariates contributing to the dynamics of judgments of misinformation are at least weakly stationary, for otherwise the equation could not hold with non-trivial predictive accuracy over time. Given that the datasets were drawn from tweets regarding news items published around 2019 by US media outlets and the 2011 Tohoku Earthquake and Tsunami in Japan, there is little reason to think that the functional form of the wave-equation generalizes to informational contexts beyond this highly idiosyncratic test environment. Hence, while it is possible that this wave-equation captures "an essential characteristic of the spread of fake news" (Murayama et al. 2021, 2), it is implausible that the kinds of relevant covariates are similar in each instance in which the classifier categorizes those covariates as fake news between some initial period and some far later time period. This is because it is plausible that newer information will come out about US media items and Japanese natural disaster events, respectively, that will lead to different features of these tweets being regarded as relevant to their classification (or lack thereof) as misinformation. In addition, there is little reason to suppose that other kinds of fake news events that have nothing to do with US media outlets or Japanese earthquakes will have similar features of misinformation. If these concerns are right, then we have a case where the probability of the covariates conditional on the labels is distinct over time, even though the independent frequency of what the algorithm considers misinformation (i.e., the unconditional probability of the label distribution) may be approximately similar. This allows us to charitably interpret the functional form of the model, insofar as it is possible that the wave equation captures some general features of the dynamics of the distribution of judgments of misinformation, while recognizing covariate shifts due to the concerns raised.

4.4 Label shift

Label shifts (LS) occur whenever a classifier considers the same set of covariates to be relevant to outcomes, but that the extent of each covariate's contribution to the judgment of misinformation by the classifier differs from training and test data. Horne et al. (2023) trained and analyzed the results of three MMMs designed as classifiers discerning whether news articles are from 'reliable' or 'unreliable' news outlets: CSN (a supervised MMM that studies the graph-theoretic structure of reliable versus unreliable news networks), NELA (a supervised text-based MMM using information from article text and headlines typically employing Random Forest methods), and BERT (Bidirectional Encoder Representations from Transformers, commonly used as a natural language processing MMM). These three models often agreed on which covariates were relevant but often diverged in terms of labels on test data, with NELA considering 86% of news sources from the Russian state-owned Sputnik News to be reliable (i.e., not misinformation), despite many international observers agreeing that this news organization often propagates disinformation.

4.5 Concept shift

Concept shift (CS) occurs whenever the probability of labels on covariates diverges between training and test data but the unconditional probability of the covariates being present is approximately the same. Consider for example distribution shifts in the context of measures of the adequacy of purely synthetic images created from Generative Adversarial Networks (GANs), a popular class of machine learning models used to generate fake images for both recreational and nefarious purposes. One metric of empirical adequacy for some MMMs, developed with GANs used in training, is the Inception Score (IS): a quantitative measure of the extent at which a GAN produces synthetically generated images that are considered indistinguishable from input images, as measured by entropy (Salimans et al. 2016).

To understand this, an image classifier is first constructed using a convolutional neural network, where the algorithm constructs a probability distribution $p(y|x)$ identifying a set of input images x with classification classes y (e.g., identifying an image, in terms of its pixels, as a dog or a train etc.), where x and y are vectors. The idea here is that $p(y|x)$ should be constructed to have low entropy in the sense that images which depict definite objects should have clear classifications with sharply peaked probabilities for being a member of a particular class. Second, we construct a GAN that generates synthetic images whose loss function is structured to incentivize confusing another 'discriminator algorithm' by causing the latter to classify the GAN's synthetically generated images with equal probability as it does with the initial (possibly real and non-synthetic) input images. The generator and discriminator are considered to have completed training when at a local Nash equilibrium in a zero-sum game, as defined by their respective loss functions (Goodfellow et al. 2020). Hence, an ensemble of generated images with distinct pictorially depictive features, considered as a whole, will have a variety of images in this ensemble that are distinguishable

from one another, while the marginal distribution over all generated images $p(y) = \int (p|x = G(z))dz$ should intuitively have high entropy. Thirdly, we compute the IS score that combines these intuitions mathematically:

$$IS(p(y|x), p(y)) := exp(\mathbb{E}_x\left[KL(p(y|x)||p(y))\right])$$

where $\mathbb{E}$ is the expectation operator, $exp(x)$ is e^x with 'e' Euler's number, and KL is the Kullback–Leibler divergence. This score returns a positive real number measuring the exponential of the average amount of entropy between each individually generated synthetic image relative to the ensemble of all synthetic images. Intuitively, higher IS numbers will correspond to successful GANs insofar as high IS values will be the result of low entropy classifications by the discriminator, while simultaneously having high entropy values for the ensemble of generated images whenever such images are distinct, leading to a set of 'individually sharp' and 'pairwise distinct' images.

While GANs are unsupervised, given that input images need not be real images depicting actual scenarios, often GANs are fed input images that are considered real by human beings so as to synthetically generate fake images (i.e., no causal connection to the real world) that are nonetheless in the same style as the input images. The philosophical significance of widespread usage of semi-supervised GANs is that what counts as a real or fake image is merely the subjective judgments of members of the set of labelers training the initial classifier used to teach the GAN how to generate fake images. The ground truth operationalization of what is considered a fake image, therefore, depends entirely on what inputs are fed into the algorithm's initial classifier: different epistemic communities with diverging standards of pictorial correctness training their respective GAN networks will lead to distinct outputs generated by their GANs. Communities will, therefore, have incommensurable standards for what is considered a real or fake image, and thus disagree about what counts as visual misinformation. And such considerations will plausibly change over time quite radically for one of at least two reasons: either new information about a picture's evidential quality is obtained, thus altering judgments about image veracity from those initial judgments (e.g., one initially interprets a politician depicted in an image as having an innocuous encounter, but later determines that their meeting is actually nefarious), or communities may change their epistemic views more generally (e.g., an old image purporting to depict ghosts is now considered fake, given the community no longer believes in an ontology of ghosts). Therefore, distribution shifts for IS measures can occur, in the sense of (CS), whenever what one society considers fake visual media changes, even for approximately similar kinds of input images and thus similar sets of covariates.

4.6 Radical shift

Lastly, as a potential example of radical shift (RS), consider the way in which the Chinese government currently regulates the development of several chatbots involving politically sensitive information (Zhang 2024, 30–31):

> "[T]he government's strict censorship policies add complexity to this situation...For Chinese AI firms, using uncensored data complicates the...process, as it increases the risk of generating politically misaligned content...[T]raining of large language model[s] often involves the use of multiple foreign open-source datasets. For example, Baidu's Erie uses primarily Chinese language data and English databases like Wikipedia and Reddit."

While the focus here is on chatbot development, the context of contemporary Chinese information politics dictates that such bots must be designed to filter and monitor what counts as misinformation, and thus *de facto* function as a species of MMM. Though there is no publicly available information on the extent at which training and test data diverge, it is plausible that there will be significant distribution shifts when the training data is mixing samples from two or more foreign datasets, such that a LLM trained with both Chinese and English databases may not have empirical adequacy when applied to real-world Chinese language informational landscapes. Here we have a potential case of (RS) insofar as English language data will inhibit the ability of the classifier to make intended predictions of misinformation judgments once it is combined with the Chinese data sets. Any automated judgments by an MMM trained on such data need not have success when English language data is combined with this set (Cf. Horne et al. 2020b). A distribution shift of this nature is even more likely to occur given the fact that what is considered misinformation by the Chinese government is idiosyncratic to that society's informational standards, as compared to most English language speaking countries' informational standards. More precisely, given the exogenous causal influence of a rapidly changing set of judgments as to what is considered misinformation by the contemporary Chinese government, especially in the wake of the 2020 National Security Law in Hong Kong, it is plausible that the presence of covariates changes over time (i.e., $P_{train}(\mathbf{x}) \triangle P_{test}(\mathbf{x})$), the set of relevant covariates contributing to judgments of misinformation changes over time (i.e., $P_{train}(\mathbf{x}|Y) \triangle P_{test}(\mathbf{x}|Y)$), and that the way in which these covariates connect to judgments of misinformation also changes considerably over time (i.e., $P_{train}(Y|\mathbf{x}) \triangle P_{test}(Y|\mathbf{x})$). This exhibits the more radical form of distribution shift that can threaten the empirical adequacy of MMMs.

5 Three potential solutions

Given these empirical case studies, I discuss three potential solutions to amend the problem of distribution shifts. Recall that MMMs are designed to automate what otherwise would be the manual informational judgments of a set of ideal epistemic elites with respect to the violation of informational preferences of members of that community. The problem of distribution shifts is that the informational preferences of a community are not stable over time and often change considerably in at least five different ways. Hence, what ought to count as an adequate solution is, at the very least, a means of rendering weakly stationary the informational preferences of those making the judgments of misinformation that will eventually be operationalized into a given MMM. Weak stationarity is methodologically desirable because it would allow social scientists to construct MMMs that can estimate the mean or variance of the distribution of judgments of misinformation using finite representative samples, enabling potential near-term policy making and the development of causal interventions to improve information ecosystems' information quality. Nonetheless, I will argue that judgments of misinformation possess a unique form of non-stationarity that ultimately renders the social science of MMMs weak, given the role that both epistemic and non-epistemic values play in what counts as misinformation. This intrinsically confounds the predictive accuracy and explanatory power of MMMs in most real-world applications.

5.1 Larger static training datasets

Sometimes an inference problem in machine learning can drastically improve by having a vast enough training dataset, as in the case of LLMs such as 'generative pre-trained transformer models' (GPTs) (Milliére and Buckner 2024). While LLMs are often fed new data dynamically in batches and via prompt engineering, it remains the case that a significant component of their learning task can be achieved using large but static data sets that do not require constantly updating datasets. In the case of some LLMs, the adequacy conditions of an LLM pertain to whether it can produce outputs which are syntactically and semantically convincing to human beings. This task is well defined insofar the syntactical and semantic features of a natural language typically do not change significantly over the period in which the LLM is deployed, despite changing over longer periods of time. In addition, such syntactical features are scrutable, and there is a sufficiently robust way to construct dialogues via prompt engineering that respect human erotetic norms and imperative discourse (i.e., how humans ask questions and make queries). This thereby allows computer scientists to employ methods from natural language processing to construct models employing cosine similarity metrics, among others, to create LLMs with reasonable amounts of success (Boleda 2020). While there remain methodological issues, larger static data sets have helped to mitigate losses in empirical adequacy.

However, it is unclear what the appropriate quantity and kind of data would be most relevant to ensuring empirical adequacy in the case of an MMM's ability to detect misinformation. Since the classification task is to identify whether some piece of natural language or image is misinformation, what constitutes misinformation is intrinsically value-laden in a way that requires specific kinds of data on the epistemic and non-epistemic values held by agents, and this is difficult to obtain. As argued in Sects. 1 and 2, it is not enough that an MMM can determine the semantic content of a proposition in order to determine whether it is misinformation, whereas this is a sufficient means of constructing a useful LLM. Rather, there are critical features of the information environment that are *external* to the language itself, that involve the cognitive judgments of human agents interpreting information that determine what counts as misinformation.

The problem for MMMs can be illustrated by the claim that 'COVID-19 vaccines are effective because they work 95% of the time'. Epistemic values pertaining to risk thresholds will decide whether such a claim amounts to misinformation or not. For instance, some may find the probability of the vaccine failing at 5% too high, instead opting for no more than a 1% failure rate. There is no objective fact as to what constitutes an appropriate threshold; nor can it be discerned from the semantic content of the proposition itself. Instead, epistemic communities will have to determine for themselves whether this efficacy rate is appropriate for their purposes, such as the financial cost of vaccines, and whether they will allow such statements to be disseminated online without being flagged as misinformation. To use another example, consider the claim that 'COVID-19 lockdowns are bad' is misinformation. This claim may be judged to be misinformation because such lockdowns are arguably highly effective for reducing mortality rates and thus are not bad. By way of contrast, one might judge this to be high quality information because it is true that lockdowns are effective for ensuring mental health and preserving economic output.

In either case, an MMM could not learn the relevant features of information from static datasets given that whether something is misinformation cannot be read from the semantic or syntactic features of natural language alone. It follows that drawing upon larger static data sets neither ensures that what humans consider false or misleading is stable over time nor that humans' epistemic and non-epistemic values are sufficiently stable over time to track human judgments of misinformation. Hence, larger data sets do not help us to solve

problems induced by any of the five distribution shifts, given that static data does not sufficiently provide the relevant kind information about the value judgments of users of language such that an MMM can reliably predict future judgments of misinformation. This is evident when we consider how *identical* propositions or images can be rationally considered misinformation in one context but not in other contexts.

5.2 Social engineering

A second solution is to have a government entity generate agreement on informational norms in a sufficiently stationary manner, such that judgments of misinformation can be reliably discerned and programmed into MMMs.

To understand the spirit of social engineering solutions, Schulz (2024) has recently proposed that we intentionally create economic systems that are in equilibrium: their statistical properties are typically easier to calculate, thereby allowing for mid to long-term predictions, they enable agents to understand in which direction states of the system are trending towards, and they enable agents to act on information more easily. One might, therefore, consider an analogous proposal in the case of misinformation by incentivizing actors to communicate information and make claims that are collectively considered by the community as unambiguous, true, and relevant to all participants. For instance, a government may consider introducing a mass education campaign teaching its citizens what it ought to say or do such that communication is unambiguous, citizens are taught a fairly homogeneous set of purported facts about reality ensuring stability of alethic judgments, and standards for what counts as relevant information are enforced through social sanctions and prudential norms. For example, theocratic societies such as contemporary Saudi Arabia, North Korea, and Afghanistan do this to a large extent by imposing homogeneous educational curricula and religious ideologies on their citizens, thereby hypothetically lowering the probability that there is disagreement as to judgments of misinformation within these countries.

However, there are numerous ethical concerns that present themselves when we consider a policy of generating informationally stable institutions in relative equilibrium for the purposes of rendering judgments of misinformation stable enough to enhance the empirical adequacy of MMMs. Firstly, while it is granted that misinformation could be easier to track, because then divergences from informational preferences will count as judgments of misinformation, this may be *unethical* in that governments ought not police speech unless such speech incites violence. After all, policing of speech will be required in order to regulate the kinds of information disseminated between relevant societal stakeholders. This at least suggests a tension between social engineering solutions, and political philosophical principles surrounding free speech and democratic discourse.

Second, as argued by Nguyen (2021) in the context of X.com (formerly Twitter), existing social engineering practices are already harmful in various ways: metrics counting the number of 'likes' and 'followers' on X.com can and have produced perverse incentive mechanisms that generate negative externalities for users, including how high quality information is disincentivized and that attention seeking information is rewarded, in terms of engagement and advertising revenue for prominent users irrespective of that information's veracity or satisfaction of other informational preferences among relevant stakeholders. This can lead to misinformation being generated if the design of the information ecosystem is misaligned between individual informational preferences and group informational preferences, where such divergence can even occur spontaneously when individuals pursue their own self-interests.

A third social engineering approach is to employ *epistemic nudges* influencing people's judgments of misinformation so that there is convergence of agreement on what is considered misinformation, thereby making it easier to automate these judgments into MMMs. The thought is that we can use psychological knowledge about people's extant cognitive biases to construct interventions, such as modifying existing user interface designs for social media platforms, that exploit these biases so as to increase the probability that users have certain kinds of beliefs, thus leading to equilibrium judgments of what count as misinformation. For example, we might try to entice people to read more about the science of COVID-19 vaccine safety by having pleasant looking advertisements placed on social media feeds which are not mandatory to look at, nor click on, but whose comparative prevalence and positive esthetic is designed to increase the probability that users click on them. The benefit of this approach is that it is a form of libertarian paternalism: it is libertarian because it allows information consumers to make their own choices without coercion, and yet paternalistic because an exogenous force is influencing aspects of consumers' belief formation processes (Miyazono 2023). The thought is that if we had enough of the right kind of data and could intervene appropriately on information ecosystems through censorship, epistemic nudges, and education campaigns, that we could empirically test and implement the efficacy of various information policies that would mitigate the spread of misinformation.

Relatedly, one might consider an intervention of *boosting* (Hertwig & Grune-Yanoff 2017). This intervention policy begins by assuming that humans are boundedly rational agents who makes decisions with limited information and have to satisfy their preferences short of their ideal. Relative to a particular agent's goals, boosting, therefore, is any intervention that a policymaker makes that enhances

the probability that the target agent satisfies their goals via knowledge of the heuristics human agents typically employ in decision making. For example, Grune-Yanoff and Hertwig (2016) advocate representing numerical information, concerning risks of using certain products, in formats that are more easily intelligible to consumers, such as representing probabilities as frequencies rather than either percentages or fractions, given empirical evidence suggesting efficacy. This method has the benefit of counter-acting what are often pernicious and yet legal methods that advertising companies employ to exploit underlying problematic heuristics and biases present in most human cognition to entice purchases. Boosting interventions which append existing informational formats (e.g., images or captions under images) can, therefore, rectify the dead-weight loss accrued from the negative informational externalities induced by advertising while preserving consumer autonomy.

However, there remains considerable practical difficulty in implementing a set of epistemic nudges that could sufficiently generate equilibrium information norms that would lead to stability in the labeling procedure for training an epistemic community's MMM. Nudges face the challenge of policymaker bias: a group of social scientists and government officials need to be in the right epistemic position to know what nudges will lead to more informational well-being than the status quo. And yet, (a) this group will be subjected to their own biases which can risk projecting their values onto consumers in a way that diverges from consumer informational preferences; (b) this group is not plausibly in the right epistemic position to be able to curate information ecosystems to a degree that it would lead to more positive than negative harms.

Regarding (a), Yee (2023a) has argued that the social media companies which run digital platforms have their own political biases that problematize automated censorship interventions that attempt to regulate misinformation flow. For example, the Myanmar government used Facebook in a manner that severely reinforced, rather than challenged, social media platform participants' views of anti-Muslim rhetoric, especially towards the Rohingya minority group that remains oppressed to this day (Fink 2018). This has led to not merely violence but even many deaths in this country, illustrating how social engineering policies related to misinformation can be disastrous when executed poorly or by nefarious government entities.

Regarding (b), Yee (2023b) has noted how MMMs are often trained in automated information environments with significant volatility and complexity, such as classifiers trained from natural language and images on social media platforms. Given the heterogeneity and complexity of these platforms, it is implausible that social scientists can construct fast enough data analysis and implement corresponding interventions at the speed and of the right kind required for effective intervention. The reason is that information ecosystems, and judgments of misinformation operative in them, are not systems whose tractability is on par with even the complexity of standard economic systems for basic consumer goods (e.g., food). In order for these kinds of social engineering policies to be successful on both a mass scale and in a manner that is actually effective in real-world applications of MMMs, one would need to provide interventions regimented with enough information on the informational preferences of consumers. But as a long line of economists and social scientists have argued, this is requires confronting an analogue of the 'Socialist Calculation Debate': the extent at which governments can design institutions that can unilaterally determine prices for economic goods and services in a manner more efficient than the free market (Boettke et al. 2024). This task is already exceedingly difficult enough to achieve for as elementary economic systems as food markets, given that Sen (1981) has argued that three of the worst famines in the 20th-century (1943 India, 1972 Ethiopia, and 1974 Bangladesh) were caused not by a lack of food but a lack of the relevant information for governments and relevant stakeholders to ensure corresponding *political entitlements* to food access were allocated appropriately, so as to prevent these famines. It is, therefore, unlikely that one could construct effective social engineering policies for far more abstract goods and services such as information on social media. The reason is that information is not a tangible object whose intrinsic value is as comparatively easy to estimate and verify as a chair or a car. For instance, it could be highly valuable information for a family to know what the weather is for the upcoming weekend, given they are planning a vacation. Given the value-ladenness of judgments of misinformation, what counts as sufficiently false or misleading information regarding the weather will be difficult to discern unless this family makes explicitly clear their informational preferences (e.g., Is it misinformation if a smartphone app displays the weather with a 30% chance of rain when it is really more like 45% chance of rain?). The point is that social scientists would need to estimate exceedingly opaque cognitive states of consumers via little to no data about their informational preferences, absent real-time survey methods.

One may object that all that matters is that MMMs can be appended by policy interventions that are *better* than the status quo, and that some social engineering interventions could achieve this. However, there are empirical and historical reasons to be doubtful of such efficacy: governments have continued to abuse social engineering methods in countries as politically diverse as the United States, China, Myanmar, and Singapore. While there may be clearer cases in which a nudge or a boost can help with certain kinds of misinformation (e.g., conspiracies about Barack Obama's citizenship and birth place), it is not enough for there to be clearer facts

of the matter about the truth-value or relevance of information such that policymakers are more justified in intervening to either reduce or prohibit certain forms of misinformation from being disseminated. An overwhelming majority of claims that are judged as misinformation will have some component of it that some informational consumer will find useful for achieving their goals in a manner that cannot be unilaterally criticized as irrational, whether by governments, social scientists, or other citizens. For instance, even the claim that 'COVID-19 came from a Wuhan lab' is underdetermined with respect to current science and thus it cannot be categorically considered misinformation independent of the context of evaluating the epistemic and non-epistemic values that feature in informational consumers' judgments of misinformation. In this sense, social engineering solutions that attempt to solve problems of misinformation remain practically difficult to implement and at worst, highly unethical insofar as they merely reinforce the informational biases of the epistemic elites who implement these interventions.

5.3 Dynamic sampling

The idea of dynamic sampling is straightforward: to enhance the adequacy of any MMM classifier, continually sample over time the informational preferences of relevant stakeholders so as to update the classifier. While the practical details are not simple to implement (Horne et al. 2019), one approach is to structure MMMs to function similarly to recommender systems, where classification is done via a softmax function as in Youtube's algorithm (Covinginton et al. 2016). In analogy with YouTube, one might create a platform in which one inputs a variety of news sources that one considers reliable and then an MMM extracts relevant features to construct a model of the user's informational preferences given user interactions with a set of news sources. For example, users could implicitly train an MMM by flagging some informational sources as 'misinformation' when they click on a button rating that source. In doing so, the algorithm could suggest other novel sources of news that are aligned with user preferences while filtering out sources that do not align, by monitoring one's interactions and dynamically sampling user behavior on the platform. Data from user interaction with an MMM can be supplemented with social scientists conducting surveys on random samples of citizens. In doing so, the relevant kinds of covariates and their relationship to judgments of misinformation can be more easily discerned as a user dynamically feeds platform data to an MMM, thereby enabling social scientists to track distribution shifts more easily. This has been the broad strategy of at least two MMMs, the fake news web application from Bojjireddy et al. (2021) and the Hoaxy web application from Shao et al. (2018); and yet, this solution has otherwise hardly been explored.

However, the philosophical conclusion to draw from this solution is that dynamic sampling, therefore, renders these kind of MMMs a species of *recommender systems*: this solution provides no prospects for a unique class of algorithms that can detect 'fake news' or 'false or misleading information' in any different manner than existing machine learning systems can recommend one's preferred music genre or shopping items on a social media platform. The grand project of constructing MMMs is, therefore, at best a cybernetic system of information preference satisfaction that is highly relative to the epistemic community in which one inhabits. This is especially so when a recent international survey ($n = 323$) of industry experts suggests the vast majority of respondents believe that current MMMs on offer are *inadequate* at detecting false information, and much better at generating disinformation than identifying it (Cassar 2023, 16). This empirical finding forces upon us a new perspective from which to design and view better algorithms produced thus far in the field of MMM research: they are value-laden recommender systems that filter and reflect our epistemic and non-epistemic preferences when making judgments about information quality. This perspective is quite different than the mainstream view according to which MMMs detect some objectively false or misleading set of propositions in natural language (Caled and Silva 2022, 125–127), or the ambitious views of social scientists of misinformation who continue to proclaim that "misinformation can be identified by both humans and machines with considerable accuracy" (Lewandowsky et al. 2024, 7). In fact, this latter claim is empirically false, as recent MMM theorists have noted (Horne et al. 2023) suggesting that mainstream misinformation researchers in philosophy and social psychology have not sufficiently engaged with the MMM scholarly community's research. It is only by paying close attention to what is required to construct an MMM that retains empirical adequacy over time that one sees more clearly the methodological issues at play.

6 Conclusion

I have argued that there is empirical evidence and theoretical motivation for the existence of five kinds of distribution shifts that compromise the empirical adequacy of MMMs. Dynamic sampling is the most promising means we have to enhance the empirical adequacy of MMMs in an ethical manner, given that social engineering has historically led to politically problematic outcomes. However, such a solution is weak in that it does not lead us to automate misinformation detection beyond that of functioning as a recommendation system for information preference satisfaction,

given the idiosyncratic ontology of misinformation and its intrinsic value-laden features. Since most societies' citizens dynamically change their values over time, and such values are not detectable merely from the natural language semantics and syntax of the static data sets that MMMs are trained on, MMMs cannot avoid the problem of distribution shifts absent some dynamically updating method. Building MMMs in this fashion renders them a form of recommender system that requires a radical change in perspective as to the nature of such systems: such systems are not literal falsity detectors, nor capable of detecting objectively 'fake news', but are rather an extension of the value-laden manual judgments of the human beings and their background epistemic communities from which such judgments must ultimately come from. It is only by being sensitive to these methodological features of MMM construction that we can begin to develop more empirically adequate MMMs that can reduce the negative impacts of distribution shifts.

Funding Open Access Publishing Support Fund provided by Lingnan University.

Data Availability Not applicable.

References

Altay S, Berriche M, Acerbi A (2023) Misinformation on misinformation: conceptual and methodological challenges. Soc Media Soc 9(1):1–13

Altay S, Berriche M, Heuer H, Farkas J, Rathje S (2023) A survey of expert views on misinformation: definitions, determinants, solutions, and future of the field. Harvard Kennedy Sch Misinform Rev 4(4):1–34

Altiabi SA (2022) Saudi Arabia's media elite's vision of the role of artificial intelligence technologies in combating social media fake news. Int J Media Mass Commun 4(2):1–41

Boettke P, Candela RA, Truitt TL (2024) The socialist calculation debate. Cambridge University Press, Cambridge

Bojjireddy S, Chun SA, Geller J. 2021. Machine learning approach to detect fake news, misinformation in COVID-19 pandemic. In: DG.O'21: DG.O2021: The 22nd annual international conference on digital government research. pp 575-578

Boleda G (2020) Distributional semantics and linguistic theory. Ann Rev Linguist 6:213–234

Caled D, Silva MJ (2022) Digital media and misinformation: an outlook on multidisciplinary strategies against manipulation. J Comput Soc Sci 5:123–159

Cassar D (2023) The misinformation threat: a techno-governance approach for curbing the fake news of tomorrow. Digit Gov Res Pract 4(4.24):1–28

Covinginton P, Adams J, Sargin R (2016) Deep neural networks for youtube recommendations. In: RecSys '16: Proceedings of the 10th ACM conference on recommender systems, pp 191-198

de Ridder J (2021) What's so bad about misinformation? Inquiry. https://doi.org/10.1080/0020174X.2021.2002187

Fathaigh R, Helberger N, Appelman N (2021) The perils of legally defining disinformation. Internet Policy Rev 10(4):1–25

Fink C (2018) Dangerous speech, anti-muslim violence, and facebook in Myanmar. J Int Aff 71(15):43–52

Floridi L (2011) Philosophy of Information. Oxford University Press

Gama J, Žliobaitė I, Bifet A et al (2014) A survey on concept drift adaptation. ACM Comput Surv 46(4.44):1–37

Goodfellow I, Pouget-Abadie J, Mirza M, Bing X, Warde-Farley D, Ozair S, Courville A, Bengio Y (2020) Generative adversarial networks. Commun ACM 63(11):139–144

Groh M, Epstein Z, Firestone C, Picard R (2022) Deepfake detection by human crowds, machines, and machine-informed crowds. Proc Natl Acad Sci 119(1):1–11

Grune-Yanoff T, Hertwig R (2016) Nudge versus boost: How coherent are policy and theory? Minds Mach 26:149–183

Guess A, Nagler J, Tucker J (2019) Less than you think: prevalence and predictors of fake news dissemination on facebook. Sci Adv 5(1):1–8

Habgood-Coote J (2019) Stop talking about fake news! Inquiry 62(9–10):1033–1065

Hamilton J (1994) Time Series Analysis. Princeton University Press, Princeton

Harris K (2024) AI or your lying eyes: some shortcomings of artificially intelligent Deepfake detectors. Philosophy Technol 37(7):1–19

Horne BD, Gruppi M., Adali S (2020b) Do all good actors look the same? Exploring news veracity detection across the U.S. and The U.K. association for the advancement of artificial intelligence, pp 1-4. https://arxiv.org/pdf/2006.01211.pdf. Accessed 16 Feb 2024

Horne BD, Nevo D, Smith SL (2023) Ethical and safety considerations in automated fake news detection. Behav Inf Technol, pp 1–22

Horne B, Nørregaard J, Adali S (2019) Robust fake news detection over time and attack. ACM Trans Intell Syst Technol (TIST) 11(1):1–23

Horne BD, Nevo D, Adali S, Manikonda L, Arrington C (2020) Tailoring heuristics and timing AI interventions for supporting news veracity assessments. Comput Hum Behav Rep 2(10043):1–16

Khan JY, Khondaker M, Afroz S, Uddin G, Iqbal A (2021) A benchmark study of machine learning models for online fake news detection. Mach Learn Appl 4(100032):1–12

Lee S, Xiong A, Seo H, Lee D (2023) "Fact-checking'' fact checkers: a data-driven approach. Harvard Kennedy Sch Misinform Rev 4(5):1–22

Lewandowsky S, Ecker U, Cook J et al (2024) Liars know they are lying: differentiating disinformation from disagreement. Hum Soc Sci Commun 11(986):1–14

Lumvembe A, Li W, Li S, Liu F, Guiqiong X (2023) Dual emotion based fake news detection: a deep attention-weight update approach. Inf Process Manag 60(4):1–20

Milliére R, Cameron B (2024) A philosophical introduction to language models part i: continuity with classic debates. https://arxiv.org/pdf/2401.03910.pdf. Accessed 24 Feb 2024

Miyazono K (2023) Epistemic Libertarian Paternalism. Erkenntnis. https://doi.org/10.1007/s10670-023-00664-9

Murayama T, Wakamiya S, Eiji A, Ryota K (2021) Modeling the spread of fake news on twitter. PLoS ONE 164:1–16
Nguyen CT (2021) How Twitter gamifies communication. In: Lackey J (ed) Appears in applied epistemology. Oxford University Press, pp 410–436
Osman M, Adams Z, Meder B (2022) People's understanding of the concept of misinformation. J Risk Res 25(10):1239–1258
Rafiqul IM, Liu S, Wang X, Guandong X (2020) Deep learning for misinformation detection on online social networks: a survey and new perspectives. Soc Netw Anal Min 10(82):1–20
Republic of Singapore (2019) Protection from online falsehoods and manipulation Act 2019. Singapore Statutes Online. https://sso.agc.gov.sg/Act /POFMA2019?TransactionDate=20191001235959. Accessed 4 June 4
Rigg B (2023) Japan's Holocaust: history of imperial Japan's mass murder and rape during World War II. Knox Press
Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X (2016). Improved techniques for training GANs. In: NIPS'16: Proceedings of the 30th international conference on neural information processing systems, pp 2234-2242
Schulz AW (2024) Equilibrium modeling in economics: a design-based defense. J Econ Methodol 31(1):36–53
Sen A (1981) Poverty and famines. Oxford University Press, Oxford
Shao C, Hui P-M, Wang L, Jiang X, Flammini A, Menczer F, Ciampaglia GL (2018) Anatomy of an online misinformation network. PLoS ONE 13(4):1–23
Søe SO (2018) Algorithmic detection of misinformation and disinformation: Gricean perspectives. J Document 74(2):309–332
Stepanova N, Ross B (2013) Temporal generalizability in multimodal misinformation detection. In: Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP, pp 76–88
Swire-Thompson B, Lazer D (2020) Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 41:433–451
Touahri I, Mazroui A (2024) Survey of machine learning techniques for Arabic fake news detection. Artif Intell Rev 57(157):1–33
Yee AK (2025) Construct validity in automated counterterrorism analysis. Philos Sci 91(1):1–25
Yee AK (2023a) Information deprivation and democratic engagement. Philos Sci 90(5):1–10
Yee AK (2023b) Machine learning, misinformation, and citizen science. Eur J Philos Sci 13(56):1–24
Zhang A (2024) The Promise and Perils of China's Regulation of Artificial Intelligence. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4708676 Accessed 29 Jan 2024