European Journal for Philosophy of Science (2023) 13:56
https://doi.org/10.1007/s13194-023-00558-1

PAPER IN THE PHILOSOPHY OF THE SOCIAL SCIENCES AND HUMANITIES

Machine learning, misinformation, and citizen science

Adrian K. Yee[1]

Received: 24 April 2023 / Accepted: 19 October 2023

Abstract
Current methods of operationalizing concepts of misinformation in machine learning are often problematic given idiosyncrasies in their success conditions compared to other models employed in the natural and social sciences. The intrinsic value-ladenness of misinformation and the dynamic relationship between citizens' and social scientists' concepts of misinformation jointly suggest that both the construct legitimacy and the construct validity of these models needs to be assessed via more democratic criteria than has previously been recognized.

Keywords Misinformation · Machine learning · Citizen science · Social Epistemology · Measurement · Construct validity

1 Introduction

Machine learning models of misinformation (MMMs) that identify and censor misinformation are increasingly prevalent in private industry, government, and academic research, including computer science (Khan et al., 2021; Shu & Liu, 2019), political science (Guess et al., 2019), engineering (Caled & Silva 2022), climate physics (Coan et al., 2021), medicine (Du et al., 2021), and especially information science (Nevo and Horne, 2022; Gruppi et al., 2021). This is so given that misinformation has spread faster and more widely than ever before due to the advent of the internet and social media platforms (Vosoughi et al., 2018). MMMs have become sophisticated and established enough that there are textbooks on core techniques (Shu & Liu, 2019) and are increasingly predictively powerful, performing well on standard performance criteria such as accuracy, precision, and recall (Mishra et al., 2022, 13; Khan et al.,

✉ Adrian K. Yee
adrianyee@ln.edu.hk

[1] Department of Philosophy, Hong Kong Catastrophic Risk Centre, Lingnan University, Tuen Mun, Hong Kong

Published online: 22 November 2023

2021, 8). This suggests that there is preliminary evidence for the empirical adequacy of MMMs.

However, our understanding of misinformation remains inchoate given that there remains considerable debate as to how to define and operationalize misinformation and cognate concepts such as 'fake news' (Habgood-Coote, 2019; Pennycook & Rand, 2021), persistent over reliance on epistemic elites' biased first-order judgments of what counts as misinformation (Yee, 2023), and the extent at which extant MMMs can capture violations of Gricean maxims in natural language processing (Søe, 2018). This raises concerns about the initial construct legitimacy of the judgments of misinformation used in extant MMMs as well as the construct validity of the operationalizations employed.

This paper discusses the epistemic workflow of MMMs, and connects this to a broader philosophical discussion in the literature on construct legitimacy and construct validity in the philosophy of science. Most MMMs typically rely upon supervised learning via the labeling of data by a diversity of stakeholders that include researchers, journalists, and average citizens recruited from services such as Amazon Mechanical Turk (Sorokin & Forsyth, 2008). I argue that this renders the construction of MMMs relevant to theories of citizen science in unexpected ways that impact MMMs' construct legitimacy and construct validity, and that this connection is a critically neglected aspect of MMMs. As I will show, extant MMMs often end up in practice overly privileging the judgments of epistemic elites who operationalize concepts of misinformation into these models. Given that misinformation is an intrinsically value-laden concept, whose operationalization necessarily concerns a diversity of stakeholders, MMMs require more stakeholder engagement than is currently the case. Furthermore, misinformation cannot be directly observed but is rather a projection onto observable information (e.g. tweets) of information quality. This renders the epistemic and metrological foundations of MMMs similar in relevant ways to discussions in the foundations of psychometrics.

This paper will proceed as follows. Section 2 begins by outlining the methods of representative MMMs while retaining technical details to the bare minimum necessary to understand the core methods. Section 3 outlines the distinction between construct legitimacy and the closely related notion of construct validity, and illustrates how contemporary MMMs are not straightforwardly construct legitimate nor construct valid, as assessed against several criteria I defend. Section 4 argues for how construct legitimacy and construct validity connect to the citizen science elements of MMMs and suggests how we might improve upon this epistemic workflow.

2 Epistemic workflow in MMMs

Nearly all misinformation scholars, including MMM theorists, tend to define misinformation as false information (Dretske, 1983, 57; Islam et al., 2020, 81; Ridder, 2022, 2). In addition to misinformation being false, some scholars acknowledge the extent at which social epistemological factors play a role in information quality: Fallis & Mathieson (2019) claim that fake news is best understood as counterfeit news that gives the misleading impression of information generated by reliable epistemic

processes typically associated with mainstream news outlets. Nonetheless, most agree with Floridi (2011) that "[S]emantic information encapsulates 'truthfulness', so that 'true information' is simply redundant and 'false information', i.e. misinformation, is merely pseudo-information" (82). Others allow for a disjunctive definition allowing misinformation to be either false or misleading information (Caled & Silva, 2022, 126-127). However, scholars have recently argued that the concept of misinformation is intrinsically value-laden and thus a function of the informational preferences of relevant stakeholders, thereby calling into question the extent at which misinformation is either purely, or even primarily, a matter of the truth-value of information (Habgood-Coote, 2019; Yee, 2023). As will become clearer, the value-ladenness of MMMs is revealed via the construction of machine classifiers, especially in supervised learning contexts. This section summarizes the construction of classifiers used in four representative MMMs, with a focus on the way in which concepts of misinformation are typically operationalized.

For instance, Shao et al. (2018) propose a network model analyzing tweets on Twitter in the time period shortly before and after the 2016 US presidential election cycle. An open software platform Hoaxy was created and used to study the dynamics of the spread of misinformation vis-a-vis engagement with fact-checkers responding to that misinformation. Hoaxy's user interface is structured to draw a representative set of tweets from Twitter's application programming interface (API) mentioning specific events using a query database search bar. The interface allows users to see how many tweets reference a particular event and allows users to visualize the dissemination of a hyperlink about that event over time via animated graphs. Articles are color coded, where grey colors signify articles of low quality and yellow colors that they have been fact checked; users coded red have behavior that is 'botlike', with blue signifying 'humanlike'. The study's findings include how there is a strong core-periphery structure (i.e. a few core tweeters spread the vast majority of misinformation), how only 5.8% of collected tweets involve fact-checking content (inducing a 1:17 ratio compared to tweets labelled as misinformation), and that fact checking websites are often shared to, ironically, further promote misinformation by framing a news item as true when it is false. Misinformation therefore appears to be widespread on the internet and yet disseminated by a few key agents. Most importantly for our discussion, the annotation procedure is conducted "by relying on a list of low-credibility sources compiled by trusted third-party organizations," the latter being journalists and other researchers who the researchers hold in high regard (3). This introduces potential bias into the training of algorithms given that these researchers are not representative of all relevant stakeholders in debates about information quality (e.g. average citizens interacting with that information).

As a second example, Castillo et al. (2012) created an MMM trained to classify misinformation pertaining to rumours in the aftermath of the February 27, 2010 Chilean earthquake. They collected millions of tweets from the time of the earthquake up until March 2, 2010 in the time zone of Santiago and constructed a graph of the retweet structure of Twitter participants discussing the event. They began by manually searching relevant cases of valid news items, with seven confirmed truths and seven false rumors that function as general stories. For each of these stories, they collected hundreds of tweets placed in the following categories: 'affirms' the news item, 'denies',

'questions', and 'unrelated'. They used a manual annotation process crowdsourced via Amazon Mechanical Turk by presenting evaluators (i.e. a non-representative sample of the general public) a random selection of ten different tweets from a topic. Each evaluator was asked to provide a short summary sentence for the topic and then asked to provide a credibility level of the topic as ranked by a four element Likert scale ranging from 'almost certainly true', 'likely to be false', 'almost certainly false', to 'uncertain'. Several machine classifiers were constructed from this labeling procedure. The first classifier labels an item as 'newsworthy' as opposed to not. These classifiers were trained beginning with the annotated topics obtained from Mechanical Turk evaluators, where several features were studied that contribute to the prediction of newsworthy topics: text-only (e.g. average length of tweets, sentiment, hashtags), user features (e.g. number of followers), topic features (e.g. most frequent URL, most frequent hashtag, most frequent user mention), and propagation retweets (e.g. the fraction of retweets versus total number of tweets). A second classification task consisted of establishing credibility scores measuring their information quality. Summarizing their workflow procedure, tweets were collected through Twitter's API, manually organized and annotated into topics by Mechanical Turk evaluators, a classifier is trained on the dataset that discerns the newsworthy from non-newsworthy tweets, and then a second classifier finds the credible topics among this set of newsworthy items. Most salient for our discussion is that Mechanical Turk evaluators, who are merely paid, non-representative members of the general public, were chosen to be the ultimate arbiters of what is considered misinformation in training this MMM. This contrasts with the previous model which consisted of epistemic elites as opposed to a sample of lay people.

As a third example, Coan et al. (2021) provide an MMM classifying climate change denial rhetoric of the past 20 years as expressed in the media content of conservative think tank communications, fossil fuel industry press engagements, and social media platforms, collecting over 250,000 blog and media posts from more than 50 denialist outlets. Firstly, research team members collected a set of statements on climate change denial from a list of myths collated on skepticalscience.com and manually categorized them into five themes, such as 'global warming is not happening' and 'human greenhouse gases are not causing global warming'. Secondly, they used a sample of 30 climate literate volunteers who were "members of a team who develop and curate scientific content on the SkepticalScience.com website" (7). However, the authors did not provide justification as to why those associated with this website ought to be considered sufficient experts to help train the machine classifiers used in their model. Thirdly, each volunteer was required to watch a video briefing them on the classifier's purpose of predicting future climate denialist claims and that the volunteers' role is to annotate texts into key themes as preparation for the training of the classifier. This example illustrates how there are often unclear criteria employed in the construction of MMMs regarding who should be responsible for labeling data in supervised contexts.

To use a last example, Jin et al. (2017) develop an image recognition classifier for identifying fake or doctored images. They collected tens of thousands of posts and images from the Chinese social media microblogging platform Sina Weibo, a platform whose structure and user interface is reminiscent of Twitter's. Weibo was chosen

because an estimated one out of three Weibo posts contains 'fake information', with an estimated eleven times as many posts with images, compared to those without images, being informationally deficient in some sense. They argue that the image-to-text ratio of real news is vastly higher than in the case of fake news, and further analyzed a variety of features of images such as "visual clarity, diversity and coherence features in a news event...resolution and popularity" (599). An ensemble of standard classifier methods were then used to annotate images by 'authoritative sources' collected from a combination of Sina Weibo itself and the Xinhua News Agency, the latter being an organization the authors allege is the "the official and most authoritative news agency in China" (601). Given that Xinhua is the official state news agency of the People's Republic of China, this example illustrates how government approved organizations can have significant influence when adjudicating information quality in the construction of MMM classifiers.

This completes our overview of epistemic workflow procedures that are representative of most contemporary MMMs. Some meta-analyses of MMMs have suggested models like these are highly empirically adequate as judged by common performance metrics in machine learning, with scores often above 90% (Alenezi & Alqenaei, 2021, 13). This is ostensibly impressive and suggests that the constructs employed at the level of operationalizing misinformation are able to track what they were constructed to measure. However, it is far from clear that this is actually the case and unclear whether we ought to accept the judgments of misinformation made by these annotators given the socially constructed and value-laden nature of misinformation. What is particularly significant about these examples is that there are typically four relevant stakeholders solicited to annotate training data: epistemic elites (e.g. trained journalists, university researchers, etc.), crowdsourced members of the general public, private, for-profit corporations like NewsScan, and government organizations. And yet, it remains unclear what criteria should be satisfied for a person or group to be considered an adequate annotator, unclear what counts as a representative sample of relevant annotators, and unclear the extent at which MMMs simply reinforce biases made by incomplete samples of relevant stakeholders.

Given that some MMM theorists have commented that most classifiers "have not reached a sufficiently high maturity level to operate without human supervision," and that "[m]any of the news veracity assessments do not accompany supporting evidence" (Caled & Silva, 2022, 143), it is important that we anaylze the metrological foundations of MMM development so that we can improve them. As I will argue in the rest of the paper, these methodological issues should not be surprising given idiosyncratic features of the dynamically updating relationship that members of the public have to the social scientists and epistemic elites who construct MMMs.

3 Construct legitimacy and construct validity in MMMs

The epistemic workflow for MMMs involves operationalizing misinformation in what are typically supervised learning contexts. As such, this raises questions as to the adequacy of such operationalizations, leading to a natural discussion of 'construct legitimacy' and 'construct validity'. The term 'construct validity' remains highly

ambiguous and is acknowledged as such by contemporary psychologists, the field in which the concept originally arose: "[C]onstruct validity continues to strike many of us, from graduate students to senior professors, as a rather nebulous or 'amorphous' concept" (John & Soto, 2007, 475). This is made all the more confusing in that construct validity and an adjacent concept 'construct legitimacy' are distinct and yet share genealogical origins in the history of psychometrics, in particular a landmark paper on educational testing by Chronbach and Meehl (1955). However, recent work has made these distinctions clear enough for our present purposes of analyzing the philosophy of MMMs.

Following Stone (2019), *construct legitimacy* can be characterized as the extent at which a construct is justified in its characterization by a theory; by way of contrast, *construct validity* is the extent at which a measurement procedure adequately operationalizes that construct. Questions concerning both construct legitimacy and construct validity typically arise when either a measurement scale is being devised in the context of a relatively novel field of inquiry or when a phenomenon is not directly observable but whose properties must be inferred indirectly via other observable phenomena. This is a pervasive method in empirical psychology, especially in the psychometrics of intelligence testing and subjective well-being studies; for instance, 'IQ tests' are measures of the psychological construct 'intelligence' (Feest, 2020).

In the context of MMMs, I will use the following taxonomy:

> **Construct legitimacy**: A concept of misinformation C employed by a group of human annotators is *construct legitimate* if and only if C is considered legitimate by the dictates of a background theory of misinformation.
> **Construct validity**: A MMM classifier M trained using a construct C is *construct valid* if and only if M adequately tracks C and produces outputs that are consonant with relevant stakeholders' informational preferences and goals.

For instance, a concept of misinformation (a construct) may be illegitimate because researchers developing an MMM might consider an item of information to be misinformation only if it is false information. As criticized by Søe (2018), this account is problematic considering that judgments of misinformation by lay people typically actually involve concepts such as misleadingness, and semantic relevance, rather than primarily truth, and that many MMMs fail to employ concepts of misinformation that are construct legitimate for this reason. For instance, a person stating truths omitting critical details misleads consumers of that information leading to deception. The only way we can assess construct legitimacy is to compare a given construct of misinformation against both psychological data, discerning how humans actually conceive of misinformation, and background philosophical theories of how we ought to conceive of misinformation, which may or may not agree with average people's concepts of misinformation. By way of contrast, as an example of construct validity, a classifier using a Naive Bayes method may do better than a different classifier using a Random Forest method at adequately operationalizing the judgments of misinformation during the supervision process.[1] Since we cannot observe misinformation directly, but rather project judgments of information quality onto observable entities like tweets,

[1] See Murphy (2022) for details on these methods.

we require classifiers to help us to automate our initial constructs of misinformation and return verdicts on novel datasets of information. However, in order to assess construct validity, it is insufficient to measure the adequacy of the construct with respect to standard performance criteria, as this would be to measure *empirical adequacy* and not construct validity. Rather, in order to assess construct validity, I will argue that what counts as adequately tracking a construct necessarily involves seeing whether the outputs from the classifier cohere with expectations from both theories of misinformation and the informational preferences of relevant stakeholders.

To see more precisely why this is the case, the epistemic workflow of typical MMMs operates[2] in the following sequence of stages:

(A) A diversity of stakeholders (e.g. journalists, average citizens, researchers, policymakers) have their own intuitive judgments about what items of information are misinformation or not.
(B) Some proper subset, typically a non-representative and small sample, of these judgments is considered legitimate enough such that the background concepts of misinformation that they employed in making those first-order judgments are considered candidates for operationalization in an MMM. I call these concepts *constructs*.
(C) Constructs are then operationalized by those training the machine classifier by labeling datasets in accordance with that construct at the level of machine code (e.g. marking specific tweets as exhibiting features of misinformation).
(D) The classifier is trained and developed via some standard algorithmic learning procedure (e.g. Naive Bayes).
(E) The classifier is fed a novel dataset and evaluated with respect to its success as measured by common statistical criteria (e.g. precision).
(F) Stakeholders then use the outputs from stage (E) to draw inferences and make decisions about the novel dataset's information quality, which may include developing government policies.
(G) There are now three possibilities. Firstly, stakeholders are happy with (E) and the process is complete. Secondly, stakeholders are unhappy with (F) and thus revise their 1st-order judgments as to what items of information ought to be considered misinformation in light of the classifier's outputs at stage (E) by returning to stage (C). Thirdly, annotators may revise their very concepts of misinformation, at the 2nd-order level, by returning to stage (B); this depends on whether the verdicts reached by the classifier at stage (F) are as expected or not.

Nearly all literature on MMMs has focused on stages (D)–(F), while stages (A)–(C), and (G) have been entirely neglected. The problem of construct legitimacy occurs during both stages (A) and (B); I further argue that the problem of construct validity occurs at stage (G) via a process I call 'cyclical calibration'. I focus in the next section of the paper on stages (A)–(C) leaving a systematic discussion of stage (G) to Section 3 of the paper.

[2] This is not to be confused with how the workflow *ought* to be constructed; as I will argue later, there are many issues with the procedure as it is typically practiced.

3.1 Construct legitimacy in MMMs

Determining the construct legitimacy of MMMs remains challenging for several reasons. Firstly, there remains considerable disagreement in the philosophical and social scientific literature as to how to define misinformation. For instance, in the MMM literature, Nevo and Horne (2022, 68) define fake news as "'intentionally', and 'verifiably' false news articles that mislead readers." Islam et al. (2020) omit the reference to verifiability and define misinformation as "a false statement to lead people astray by hiding the correct facts" (81), and Shu and Liu (2019) define 'fake news' as "a news article that is intentionally and verifiably false." In the philosophical literature, Dretske (1983, 57) wrote that "false information, misinformation...are not varieties of information - any more than a decoy duck is a duck." In contrast to these alethic views of misinformation, Swire-Thompson and Lazer (2020, 434) define health misinformation 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." Similarly, Hou et al. (2019) articulate health related misinformation as "incorrect information that contradicts current established medical understanding." This presents a relative conception of misinformation as information that is deficient relative to the highest epistemic standards of the time. Coan et al. (2021) take a different approach and define (climate change) misinformation as claims that "have been shown to contain reasoning fallacies" (3).

Secondly, despite considerable disagreement at the level of theory, it is overwhelmingly the case in practice that the construction of MMMs involves deference to epistemic elites' first-order judgments as to what is true and false during the supervised learning processes of MMM construction. However, both what the truth is and the concept of truth employed is simply taken for granted as obvious in judgments of misinformation, with deference nearly always given towards epistemic elites such as seasoned journalists, subject matter experts, and academic researchers. Though these latter groups have clear epistemic strengths, they are not necessarily the best guide to the information quality of information concerning either novel information, information which directly concerns the lived experiences of the under privileged, or information whose quality is better adjudicated by a diversity of stakeholders. While some have argued that structuring society with epistemic elites in power can satisfy several political philosophical virtues such as better policy making (Brennan, 2016), recent concerns have been expressed that MMMs and other models used to study misinformation risk exacerbating underlying biases in the informational judgments of the annotators that risk automating pernicious forms of epistocracy (Yee, 2023). Such biases are not merely epistemic but are often ethical or political. For example, machine learning could potentially be used to worsen the effects of recent legislation from governments such as Singapore, a country which has passed misinformation laws since 2019 that arguably justify undue exercise of power over journalists and activists who speak against the government, rather than mitigate misinformation in ways citizens really care about (Republic of Singapore, 2021). These controversies cannot be ignored and yet remain neglected in recent discussions of MMMs.

Despite these methodological concerns, stages (A) - (C) typically proceed in practice as follows. MMMs contain both observable entities (e.g. tweets) and unobservable entities (e.g. the semantics and information quality of the tweet). To use an example of a common method, a tweet's informational quality is measured by way of labelling a tweet as misinformation. This typically occurs when an annotator, perhaps hired via Amazon Mechanical Turk's labor supply, judges that the semantics of a tweet ought to be interpreted in a certain way that suggests a deficiency in information quality. However, information is socially constructed and context sensitive to an agent's interpretation of that data (e.g. a literal reading of a Tweet versus what it semantically implies given the context) and every agent's interpretation of that information is a function of their background community's epistemic standards in which that information is conventionally interpreted and understood (e.g. an anti-COVID lockdown tweet may be classified as misinformation in one community but not another).

Thirdly, the most common epistemic standard for evaluating information quality in the context of MMMs has been whether or not the item of information is true or not. However, it is far from clear that there is a concept of truth that MMM theorists and stakeholders can agree upon that will not contain intrinsic epistemic controversies. To see this, consider firstly that even our best epistemic practices, namely the natural and social sciences, typically do not require a concept of truth as a cognitive value as compared to the satisfaction of other cognitive values pursued in scientific inquiry, such as predictive and explanatory power, consistency, and parsimony (Elgin, 2017). This is especially so when one considers the non-trivial error terms common in any regression method (e.g. ordinary least squares) in the sciences. This error is sometimes taken to be innocuous for practical purposes, but acknowledged nonetheless as part of an explicitly false though highly useful model for a variety of instrumental purposes. More strongly, one can run a pessimistic meta-induction over the history of science and argue that most scientific theories in the past are now considered to be false, insofar as we consider them to posit entities and structures which are non-referring (e.g. phlogiston theory of chemistry, humoral theory of medicine, Darwin's gemules in biology, etc.) (Laudan, 1981). This has led constructive empiricists to claim that scientific practice ought to aim at most at satisfying standards of empirical adequacy (van Fraassen, 1980). Hence, misinformation cannot be defined as information structured as false propositions as this would entail that either most of natural and social science is misinformation, which is absurd, or that misinformation can be highly predictively and explanatorily powerful like the sciences, which abuses the term misinformation.

If one is not convinced of examples from science, the lack of clarity on the concept of truth in the context of quotidian examples of misinformation also arises. Consider how some MMMs have been trained to identify fake photographs, as in Jin et al. (2017), considering that there are estimates that more than half of posts on the popular Chinese social media app Sina Weibo contain images accompanying text. Photographs are not literally true or false, as they are not even propositions. Rather, what typically makes a photograph an item of misinformation is the extent at which it can be misleading, where misleadingness is a function of the hermeneutic conventions of an epistemic community interacting with the photograph. The confusion arises in that when one thinks a photo is fake, in the sense of not accurately depicting some state of affairs, this is because the photo is presented in a context in which the implied pictorial semantics

of the photo are intended to be interpreted as a matter of depiction, rather than, for instance, altered images for the purposes of art or as a joke. Hence, Jin et al. (2017) hypothesize that one adequate measure of the extent at which an image is fake (i.e. fails to accurately depict an event) is the extent at which an image is significantly different in features than other images taken of the same purported event. But even here, it is not clear what counts as the event in question considering that each photo is, strictly speaking, distinct: we see a different angle of a politician from one camera, there are different people shown in the event, etc. What disambiguates this underdetermination in practice are informational norms, which are culturally and epistemically contingent upon one's upbringing, societal norms, and one's education. These features are so fundamental that they are often taken for granted. And yet, diverging interpretations of the meaning or significance of the same photograph happen all the time in judgments of misinformation. Hence, it is not clear that there is a single objective depiction of an event of which any given photo can be considered more or less accurate, with respect to its depiction of, without relying upon background epistemic and informational norms which are intrinsically negotiable.

A related point has been made by Søe (2018) according to which the diversity of thresholds in which standard Gricean norms of relevance are violated has already created considerable confusion surrounding the intended application and purported success of MMMs. For instance, consider those who proclaim that 'COVID-19 vaccines are ineffective'. This claim has often been described as false and therefore misinformation, considering the relatively high efficacy of most vaccines at mitigating severe symptoms of COVID-19. However, this judgment is controversial for many reasons. Firstly, while the frequency is extremely low, some side effects have been known to occur, such as blood clotting in the Johnson & Johnson vaccine (Mahase, 2021), and some lay people consider this evidence that the vaccine is 'not effective'. In a situation such as this, some observers of these effects might protest that while the probability of accruing a side effect is very low, they nonetheless weigh the outcome of experiencing a severe side effect very highly as a negative outcome to be avoided. It follows that the expected utility (i.e the multiplicative product of the probability of an outcome's occurrence and its utility) of believing in the dangers of vaccines is enough that they may decide to make the claim that vaccines are ineffective. While there may be other reasons to resist this line of thought, it remains the case that this is arguably a consistent view to hold and one which ought to be taken seriously in any discussion of health misinformation and how to enhance trust in vaccines (Goldenberg, 2021). Hence, values play a critical role in adjudicating what counts as misinformation. If an MMM is not properly trained to incorporate value judgments in the process of annotating, then there is a lack of sufficient construct legitimacy to the concept of misinformation being employed in supervised learning contexts.

Fourthly, there are sometimes significant cases of underdetermination of theory by evidence. Consider ongoing discussions as to the origins of COVID-19, such as whether it came from nature via zoonotic transmission or from a lab leak. Many proclaim that the lab leak theory is a conspiracy theory tantamount to spreading misinformation; and yet, it remains under determined as to the disease's origins. In cases such as this, it is unproductive to protest that one simply does not agree on the assignment of credences as to the disease's origins being from a lab and that the probability

is higher that it came from nature. Other conspiracies such as the September 11, 2001 terrorist attacks are similarly either vague in their exact pronouncements or are the result of disagreements about expected utility calculations at the level of values, given that it is values which influence the quantity of utility assigned in an agent's utility function that leads them to pursue a particular line of inquiry about a conspiracy theory. Hence, a person may rationally believe in a conspiracy theory, as measured by internal consistency and expected utility calculations, that others would consider misinformation. This shows how it is not straightforward how to operationalize concepts of misinformation without recourse to judgments of information quality from a diversity of relevant stakeholders who will disagree about fundamental epistemic factors related to probability judgments about the reliability of evidence.

Fifthly, recent psychological data from Osman et al. (2022), surveying $n = 4,407$ from four countries (Russia, Turkey, UK, and US), suggests that as much as 69% consider misinformation as 'information that is intentionally designed to mislead', and that 49.24% thought that misinformation was information that typically 'exaggerated conclusions from facts', 'didn't provide a complete picture' (48.83%), and was 'presented as fact rather than opinion or rumour' (43.07%). This shows that while judgments of truth-value play a component of lay people's understanding of fake news, it is not clear that it is a necessary component, nor even the majority component of judgments of misinformation. This further suggests that factors such as the intention, epistemic relevance, or the salience and granularity of an item of information are integral to the concept of misinformation in many people's minds. MMMs have often continued to ignore these findings from the empirical psychology of misinformation and thus routinely posit controversial constructs of misinformation in supervised learning contexts.

The point here is that what conditions one's judgment that an item of information is tantamount to misinformation will involve background epistemic and value-laden assumptions that are not merely alethically oriented but which are often intrinsically up for debate given their moral or political nature. That there can be significant consensus of agreement on a particular subject matter or event's occurrence (e.g. COVID-19 vaccines are more effective than not), is therefore rarely sufficient to justify the charge that someone is objectively sharing misinformation. Information quality is a function of a given agent or informational community's interests in obtaining that information, such as whether that information is relevant, whether it is misleading, whether it is at the level of granularity that is commensurate with their interests, whether it allows them to make predictively accurate claims, whether it is explanatory, or whether it coheres with one's background cognitive and non-cognitive values more generally. Misinformation therefore ought to be understood as a *relative* term where judgments of misinformation are formulated relative to one's own or one's community's informational preferences and values.

To see more concretely how this impacts both the construct legitimacy and the construct validity of MMMs, we revisit the epistemic workflow in the Coan et al. (2021) MMM study on identifying climate change denial rhetoric. While there has been expert consensus since at least 2004 that humans have caused climate change at rates we have witnessed in modern times (Oreskes, 2004), average citizens who are stakeholders may nonetheless think that the normative implications of these findings

are unclear. Some citizens may want a quick transition to renewable energy sources that reduce fossil fuel emissions; others may disagree and think that climate activists are spreading misinformation by overstating the severity of climate change and that such a transition should either occur slowly or not at all. What counts as overstating or understating the severity will be an intrinsic function of one's background epistemic and political values. Hence, in the context of MMMs, the construct legitimacy of the judgments of misinformation being used to annotate datasets will be a function of whether one considers the annotators epistemically reasonable and competent or not. It follows that any operationalization of misinformation will incorporate the epistemic assumptions of the research team and participants constructing the MMM, despite such assumptions possibly being considered unreasonable by other perfectly rational, distinct stakeholders. This introduces a problem for evaluating the construct legitimacy of MMMs at stages (A) and (B) of the epistemic workflow considering such disagreement. Therefore, MMM researchers ought to acknowledge the controversial epistemic situation they are in, in the sense of perpetuating background epistemic biases through machine classifiers, and remain transparent about these potential weaknesses of their supervised learning procedures. This is especially so when the classifier developed by Coan et al. (2021) was trained via the annotations of a biased sample of 60 undergraduate students, none of who are representative of the full spectrum of relevant stakeholders that this classifier's outputs concern.

This completes our discussion of construct legitimacy; we now connect this discussion to the problem of construct validity in MMMs.

3.2 Cyclical calibration and construct validity of MMMs

In analogy with the situation in psychometrics, such as the development of intelligence tests, the construct validity of a MMM classifier ought to be assessed with respect to the extent at which that classifier both (a) tracks the construct as intended by producing outputs that stakeholders consider consonant with their informational preferences and (b) the classifier is consistent with the results of other MMMs that have been constructed on similar topics and with annotators of similar epistemic disposition. While there are comparatively unexplored questions regarding the empirical adequacy of MMMs that (a) raises, this goes beyond the scope of this paper; rather, it is (b) in particular that is our present focus. The reason (b) ought to be strived for is that satisfying it suggests robustness of the measure and overall coherence with other humans' judgments from similar epistemic communities on the information quality of similar informational items. Since these others will by assumption have similar epistemic dispositions, and thus be considered a member of the same broader informational community, a classifier returning similar kinds of results ought to raise confidence in the construct validity of the classifier. Nonetheless, there are additional features that require exploring as to the exact relationship that all other stakeholders have to the evaluation of construct validity, especially considering many annotators of datasets are sampled from members of the general public. Some members will be epistemically competent and others will not be. This raises questions as to what informational

preferences matter, whose preferences might matter more than others, and who counts as part of the same informational community.

To gain some traction on this issue, we note that the construct validity of MMMs is a function of the extent at which we consider epistemically reasonable those involved in the annotation process of supervised learning. We can further identify several core features of what can be described as the *cyclical calibration procedure* that occurs throughout stages (A) - (G) and back again to (A). The procedure is cyclical in that it often repeats, either in a single study (Horne, 2020) or understood as a research paradigm consisting of multiple MMMs developed on a similar topic (Caled & Silva, 2022). It is furthermore a process of calibration given that it requires continual adjustment and refinement in light of testing the classifier against prior operationalizations of informational judgments from annotators. While these and other extant studies do not explicitly acknowledge that they engage in a process of cyclical calibration, I will show how they are nonetheless arguably necessarily implicitly committed to such procedures in practice.

Firstly, social science researchers and other annotators are involved in a hermeneutic circle when they construct MMMs, given their simultaneous roles in scientific, public, and private discourses of information assessments. Given this process, disagreement is inevitable and convergence of agreement is not always the case. Here, the phenomenon to be predicted and explained (i.e. misinformation) is entirely a function of the beliefs of individuals; this is not the case in the context of phenomena, for instance, in the physical sciences, where a particle's properties will be at most a partial (and not entire) function of the contingently dominant scientific community's ontological and metrological standards.[3] Notice that this issue is not salient in the case of most natural scientific contexts such as modern physics, where the mass of an object is intersubjectively verifiable via measurements using quantities from an objectively defined SI system of units. No such intersubjective verification is possible in the case of misinformation given that informational quality is not uniformly experienced in a raw, sensory format but contingent upon the idiosyncratic informational judgments of relevant stakeholders, which are coordinated mental projections of informational quality onto items of information themselves (e.g. tweets). That is, while we may safely assume that most humans could agree on the literal words expressed by the same tweet, it does not follow as a necessary consequence that each will agree how to interpret that tweet and therefore adjudicate its information quality.

Secondly, there is a lack of consistency as to the ontology of the phenomena: what is considered misinformation changes rapidly, sometimes undermining the ability to identify future instances of misinformation in either a consistent manner with previous verdicts or in a robust fashion (where multiple independent measurement methods often diverge in their agreement as to what misinformation is). For instance, ongoing debates concerning the extent at which the claim that COVID-19 came from a lab in Wuhan illustrate how mainstream media has vacillated on the extent at which this is misinformation. In this sense, misinformation is not a stable phenomena in the way the properties of, for instance, an electron's mass are constant over time. One way

[3] See Franklin (2016, 229-240) for a discussion of how varying thresholds for statistical significance have even decided the very ontology of sub-atomic particles.

to put this point more precisely is to say that the distribution of judgments of items considered to be misinformation from annotators working to construct MMMs, is not even approximately drawn from a stationary stochastic process.[4] This makes it difficult to track concepts of misinformation in a community of stakeholders. This impacts construct validity for three central reasons: (i) ostensibly similar structural properties between datasets can be illusory and confound MMMs; (ii) the same set of users can radically change their habits of sharing misinformation given new belief formation, or due to exogenous causal factors; (iii) the same information platform can change its policies on misinformation quite drastically and suddenly, disrupting ostensible equilibrium properties of the network sampled from (e.g. Facebook spontaneously censoring and removing misinformation related to COVID-19 vaccines).

The first point (i) is highly non-trivial and has already raised issues in MMM construction. Horne et al. (2020) trained 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 purported epistemic elites from the organization Media Bias/Fact Check, their classifier methods surprisingly struggled to perform well, 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), given that US and UK English are distinct dialects. Furthermore, combining both the UK and US training data does not help to enhance success either. 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. Hence, despite ostensible structural similarities in two populations, idiosyncrasies between two dialects of a language can seriously confound MMMs' predictive powers.

The second point (ii) has also been neglected in the MMM literature so far. How a person comes to understand informational quality is critically tied to the kind of misinformation they will spread; after all, if a person does not think some information X is misinformation, then they are more likely to spread it than if they thought X was misinformation. For example, if a hypothetical person formulates an epistemic rule and considers any rhetoric by Donald Trump during the 2016 US election cycle as misinformation, then this person will be biased to decline the sharing of information disseminated from Trump on the basis of this rule (e.g. 'if proposition p is asserted by Trump then p is misinformation and should not be spread'). Furthermore, our concepts of misinformation are arguably constantly dynamically updating. To use a historical example, Russian citizens living under Stalin's government came to learn from the Smolensk Archive, first publicly published in 1958 by historian Merle Fainsod, that they were victims of systematic mass propaganda and frequent disinformation campaigns. While many Russian citizens knew that there were serious informational problems in their society, the scope and scale was not fully clear. This eventually altered many citizens' former concepts of what misinformation is and what

[4] See Brockwell and Davis (2016, 13) for a precise mathematical definition. In essence, the behavior of stationary systems have stable statistical properties in its first and second moments for any given time-lag shift of that time series.

its features are (e.g. that their government was even more nefarious and malfunctioning than they realized) (Arendt [1951], 1976, xxv).

In this sense, the very ontology of misinformation, and information more generally, is a direct function of the collective intentionality of epistemic agents' coordinated acts of regarding data amongst a landscape of competing informational judgments and preferences. This illustrates the sense in which stage (G) has three possibilities since both the 1st-order judgments of what items of information are misinformation, and concepts of misinformation at the 2nd-order level, are dynamically updating, in light of agents' changing epistemic environments. This implies that both MMMs' construct legitimacy and construct validity is contingent and often transient, rendering the present science of MMMs weak in terms of predictive and explanatory powers.

Lastly, as for which stakeholders' preferences should take precedence over others in any given period of deliberation, this cannot be decided *a priori* but must be sensitive to the specific informational goals stakeholders have at the time in local contexts of debate and discourse. This arguably requires a theory of *preferentialism* according to which an MMM is adequate overall only if it is both sufficiently construct legitimate and construct valid relative to enough relevant stakeholders' informational preferences. In analogy with work done by Alexandrova (2017, 150), and her account of the construct validity of psychological measures of subjective well-being, I argue that the ideal set of criteria that ought to be satisfied for a classifier M to be considered construct valid, given a construct of misinformation C, is the following:

(I) M is given labelled data using annotators who are sufficiently representative of the relevant stakeholders to which concepts of misinformation (C) apply in the construction of MMMs.
(II) M is consented to as much as possible by relevant stakeholders of which C directly applies.
(III) Background psychological and epistemological theories of C, and the moral and political values of stakeholders, are largely consistent with variations in M's outputs across a diversity of relevant and novel datasets that M is provided.

I provide justification for each of these three ideals (I) - (III) in the next section on the topic of the citizen science elements intrinsic to the construct validity of MMMs.

4 Citizen science and MMMs

I have argued that we require that the kinds of informational preferences of those annotating datasets in supervised learning be representative of the preferences and concepts of misinformation that will be employed by relevant stakeholders regarding the classifier's output. However, each annotation group need not necessarily have *identical* preferences or concepts given possible divergences of informational needs across each group; that they have sufficiently shared preferences or concepts is enough. This is because there are typically four relevant stakeholders in the construction of MMMs: average citizens, government policymakers, social scientists, and journalists.

In this section, I will assume for now that representativeness ought to be beholden to standards within the liberal democratic tradition where every individual stakeholder

has some means of expressing their informational preferences and can have a non-trivial probability in having those preferences satisfied. I leave it to other work to decide how cyclical calibration ought to proceed in societies that are not liberal democratic in nature.

4.1 Justification for ideal (I)

Average citizens are the primary consumers and propagators of informational discourse and hence their beliefs and concepts of misinformation are most important in constituting the ontology of misinformation. Combating misinformation is important as it can be a matter of life or death for civilians, such as in the context of misinformation surrounding purported cures for COVID-19. Furthermore, civilians form the backbone of groups who are most directly involved in issues of trust in science and lack thereof. MMMs have in fact already often relied upon average civilians who perform annotation tasks, via crowdsourcing services such as Amazon Mechanical Turk or Crowdflower, and hence bring their own informational preferences when adjudicating information quality. By way of contrast, government policymakers have different priorities in evaluating and conceiving of information in that they have an eye towards either satisfying some set of domestic or foreign policy objectives, as in the Government of Canada's efforts to combat Russian disinformation during the ongoing Ukraine conflict (Government of Canada, 2022a, 2022b) or towards improving some form of general social cohesion (Republic of Singapore, 2021). In cases such as the Republic of Singapore, misinformation is defined relative to both the truth-value of that information and the extent at which that information can be used to challenge the government and disrupt societal status quo. Hence, government policymakers' concepts of misinformation are typically structured to be oriented towards satisfying narrower goals than the diversity of consumer informational preferences and needs of the general public. Lastly, social scientists and journalists are interested in trying to study misinformation in a way that they can predict and explain misinformation, and communicate their judgments to members of the public and government stakeholders.

This illustrates why desideratum (I) is important to satisfy in MMM classifier development. In these ways, MMM model construction is constitutive of a disguised form of citizen science in which members of the public play not only a cognitive labor role in the construction of MMMs at the level of annotations but even a constitutive role in what misinformation is. Considering how high stakes the debate concerning what misinformation is, how it spreads, and its direct relevance to laypeople and academics alike, it is arguably critical that the social scientists who already often use citizens' cognitive labor come to serve their interests on equal par with the interests of citizens. As it stands, this is not the case; the current epistemic situation is therefore problematic given a lack of sufficient awareness of these methodological issues.

4.2 Justification for ideal (II)

We can improve this situation by drawing on insights from feminist epistemology, according to which scientific knowledge is more objective only if more relevant stake-

holders' viewpoints are incorporated into the epistemic workflow of scientific model development. For example, Longino (2022) recently argued for a view she calls Critical Contextual Empiricism (CCE) according to which scientific knowledge is knowledge that requires critical interaction amongst community members according to community norms of knowledge acquisition attained at specific granularities, where these granularities are decided primarily by both pragmatic concerns of stakeholders and their non-epistemic values. What makes her account 'critical' is that through sustained engagement and criticism with others within a scientific community, not only are our assumptions supported, refuted, or amended, but are also made publicly explicit in their content. Mutual deliberation therefore serves an edifying role that can assist in scientific model development and render it more objective by reducing bias.

To import CCE's insights into the context of the development of MMMs, the dynamically updating features of both laypeople and researchers alike is a testament to the features of normal science that Longino highlights, and yet have not been recognized as such in the MMM literature. After all, MMM theorists are engaged in reciprocal relationships of trust between journalists and crowdsourced citizens in unforeseen ways at the level of annotation that can bias classifiers' judgments. However, almost no dialogue between each group of stakeholders has taken place in MMM development that could very well potentially enhance the standpoint objectivity of the entire epistemic workflow procedure. Moreover, dialogue ought to be conducted in a manner that enhances the extent at which stakeholders can consent to the annotation procedure. While this need not entail consent from absolutely all members, as there will inevitably be significant disagreement, obtaining as much consent as possible ought to be the ideal.

In these ways, MMMs are often already in practice a disguised form of citizen science and yet, MMM researchers have failed to sufficiently explicitly acknowledge this, let alone notice. This naturally raises questions as to what form of citizen science is best to regulate the epistemic workflow of MMMs. While there are many views on how citizen science ought to be conducted,[5] the European Citizen Science Association's ten principles suffice for our purposes (Robinson et al., 2018, 29-30):

(1) Citizen science should generate new knowledge or understanding by having a meaningful role in knowledge production.
(2) Projects require addressing some scientific goal such as prediction or explanation.
(3) Scientists and citizens should both mutually benefit from shared research practices.
(4) Citizens can in theory participate at any, or multiple, stages of the scientific process.
(5) Citizens should receive feedback from researchers.
(6) Citizen science has limitations and should be recognized as such.
(7) Projects should publicly disclose data and metadata and ideally publish in open-access journals.
(8) Citizens' roles should be acknowledged in the final paper or report.
(9) Projects should be evaluated for their wider societal impact.

[5] See Hecker et al. (2018) for an overview of recent literature on global citizen science initiatives and their philosophies.

(10) Leaders of citizen science projects should consider legal and ethical issues surrounding data integrity and privacy, and any environmental impact of activities conducted.

Principles (1)–(4), and (7) are already typically practiced by MMM theorists to varying degrees while (5), (6), (8), (9), and (10) are often neglected. (5) and (10) are important but we put these aside for now, as they are beyond the scope of this paper, and focus on (6), (8), and (9).

(6) is relevant in that MMM theorists routinely treat their reliance on a highly heterogeneous group of annotators in supervised learning contexts as relatively uncontroversial, and are too focused on satisfying statistical performance metrics. Acknowledging limitations is important given what are sometimes severe disagreements amongst annotators. For instance, recall the Coan et al. (2021) study design and their decision to only allow annotators who are 'climate literate', thus biasing the classifier in accordance with their specific threshold for climate literacy, which may not be judged to be adequate enough by other stakeholders given their preferences and social goals. (8) is critical in that readers of MMM studies ought to be participant to the process of adjudicating the extent at which annotators are sampled from sufficiently representative stakeholder groups. Lastly, (9) suggests MMM practitioners should note the risks that classifier biases can play in the contexts of their usage, given a wide variety of actors use them for radically distinct purposes, thus exposing civilians to a diversity of epistemic risks.

To satisfy these desiderata, we might consider three proposals. Firstly, we might want our MMMs' outputs to be agreed upon by greater than 50% of each relevant stakeholder's respective group. After all, if this were not the case, that would appear to be unsatisfactory from a democratic perspective. A second option is to weigh the stakeholder groups unequally but in proportion to some metric of risk. For instance, an idealized, benevolent government may wish to weight a stakeholder group's informational preferences more heavily if that group could be the victim of genocide, given a piece of misinformation is disseminated and failed to be flagged as misinformation. A case in point is the ongoing situation in Myanmar concerning the persecution of the Rohingya ethnic minority group, given Facebook's history of condoning social media posts which incite violence against them (United Nations Human Rights Council, 2018, 165). In this sense, Facebook's algorithms for flagging misinformation are clearly inadequate if they are inattentive to the informational preferences of the most important and vulnerable stakeholders. Hence, an unequal but proportionate weighting ought to be given towards enabling relevant Rohingya people the opportunity to assist in the adjudication of information quality in MMM development. Thirdly, one could consider a weighting which is not uniform across stakeholder groups but is instead stochastic. The idea here is that we can ensure representativeness of a sample in a negative sense by removing any possibility of bias in the sampling procedure via random sampling across the whole population, and not just a subset of purported experts. It has been argued by Guerrero (2014) that having a society based on this format of 'sortition' is one means that one could satisfy democratic principles of representativeness while nonetheless sacrificing the obvious virtues of voting procedures. To amend this proposal in the context of MMMs, instead of choosing our sample of stakehold-

ers in a stochastic manner, we could instead assign our weights stochastically so that while each stakeholder is chosen non-randomly, the importance (i.e. weight) of their informational preferential contribution to the total sum of informational preferences needing to be satisfied is nonetheless stochastic. This would help to ensure that the democratic principle of representativeness is appealed to in an unbiased manner.

This is but a sketch of three methods and each account of representativeness will have its benefits and shortcomings; what procedure MMM users will adopt cannot be decided *a priori* but will be decided within the specific context of usage. What is key is that the citizen science aspects of the annotation procedure will require some account of democratic participation along these lines in order to ensure construct validity in the sense of cyclical calibration.

4.3 Justification for ideal (III)

Despite ostensible virtues of my proposal, I acknowledge that citizen science of the kind our discussion concerns can nonetheless be liable to induce significant potential problems unless care is taken. To use a hypothetical scenario to illustrate, consider the existence of a completely isolated and homogeneous community of astrologers (a field of inquiry we assume for now is epistemically deficient) who supervise an MMM to classify tweets spread in their online communities which contradict astrological findings as misinformation. Suppose further that the MMMs score very highly on a variety of statistical criteria (e.g. precision, recall, etc.) and thus are ostensibly empirically adequate. Furthermore, since this entire community is by hypothesis uniformly distributed in its informational preferences (i.e. all are believers in astrological theory), it would appear that the MMM's judgments of misinformation are not merely ostensibly but actually in fact construct valid as well. This is because it would appear to satisfy my aforementioned proposal, particularly concerning representativeness of the relevant stakeholder populations' informational preferences, even if astrology itself is an epistemically deficient theory as adjudicated on other epistemic grounds (e.g. it often makes wrong predictions). Hence, astrologers using MMMs can have MMMs which are construct valid while nonetheless flagging anti-astrological tweets as misinformation.

While this is ostensibly problematic given that astrology is widely considered now to be predictively and explanatorily deficient, I am wholly willing to accept that scenarios of this kind are a peculiar but important consequence of my theory to be acknowledged. Notice that in this example, the population has already converged in their agreement on a background epistemic view; the assumption of homogeneity of the groups' informational preferences renders it such that there are no dissenters here. One may object that this is still highly problematic given that there could be many cases in which, for instance, mainstream astronomical findings would be labelled misinformation (e.g. that a large asteroid is coming to earth that could end life on earth), and vague astrological findings will be neglected to be labelled misinformation that could be potentially life threatening. This would severely compromise other aspects of citizen well-being which appear to require equal consideration when considering the legitimacy of any epistemic ecosystem. However, even though their belief in astro-

logical theories could be *de facto* harmful to this homogeneous group of astrologers, even if they are not aware of it or do not believe, this is not a sufficient objection to my view of citizen science. The reason is that misinformation cannot be defined merely with respect to purported falsity or ignorance. Given that even the best scientific communities were considered to have been mistaken throughout history (e.g. classical mechanics radically fails to make even approximately correct predictions in the quantum realm), despite participants forming beliefs about the world with the best evidence they have, it is unreasonable to refer to a situation such as this as misinformation given there is no misleading component to the way these astrologers are sharing and processing information with one another. That is, this community had adhered to the highest epistemic standards that were possible in that community. In this sense, judgments of misinformation should be understood as relative terms and measured against the satisfaction of the informational preferences of the most number and variety of stakeholders in a society. That this community constitutes an echo chamber, in the sense of Nguyen (2020), is an independent objection one could make that is not directly relevant to the question of the construct validity of MMMs. After all, in this homogeneous community, all stakeholders believe in astrological theory and thus all relevant informational preferences are assumed to be satisfied in this community.

Moreover, we can assume that this community is one in which no one was coerced into holding astrological views, and thus the formation of informational preferences was entirely consensual, in accordance with ideal (II). It is critical in this example that this community be isolated so that the set of relevant stakeholders is homogeneous in preferences and values; for otherwise, if there were another non-astrological community that was exposed to the consequences of an MMM supervised by astrologers, then the situation would be different and an alternative verdict ought to be reached that the MMM is not sufficiently construct valid. Hence, the relevant stakeholders in the construction of an MMM are not only those who are part of the supervision process but also those who can be affected by the outcomes of the supervision process, as outlined in stage (F) in the epistemic workflow of standard MMMs.

This being said, the situation involving this homogeneously distributed astrological community ought to be evaluated differently from a closely related but distinct scenario in which *not* all stakeholders are consulted. To illustrate using historical fiction, imagine that we were in the time of former US president George Washington, who contracted an illness and was given the attempted cure of bloodletting which, as a matter of historical fact, tragically led to iatrogenically caused death. Despite it now being considered an ineffective and iatrogenically harmful treatment for most ailments, bloodletting was prescribed during Washington's time given that it was the dominant view in that period of the history of medicine (Chatham, 2008). Now suppose anachronistically that an MMM was constructed, and supervised via the annotations of prominent medical practitioners working with computational linguists in the 1700s, to scan an equally anachronistically existing internet and flag posts on social media from people proclaiming bloodletting to be ineffective. Unlike the case of the astrological community, it was actually the case that many people suffered and were often agreed to have suffered from iatrogenic causes due to bloodletting. That is, there was not completely homogeneous belief in the efficacy of bloodletting insofar as not all relevant stakeholders for whom bloodletting applied were having their informational

preferences satisfied. This is because many people died as a result of bloodletting and who believed that it was bloodletting that was the cause of their death. By way of contrast, in the hypothetical astrology case, everyone is assumed to believe in astrological theory even if their objective well-being may sometimes be compromised by their false beliefs. In the historical fiction case, it is accurate to say that the MMM was not construct valid given the failure to satisfy the relevant moral and political goals of all relevant stakeholders. That bloodletting was considered at the time to be the best theory satisfies only half the desiderata for construct validity as in ideal (III). Hence, the process of building an MMM in this counterfactual historical scenario during stage (G), among stages (A) - (G), would be considered deficient given that not all relevant stakeholders' preferences were being satisfied. This therefore fails to satisfy ideal (III) and illustrates how cyclical calibration is sometimes satisfied in populations with comparatively homogeneously distributed preferences as compared to those which are not.

We close with a comment on how citizen science is an important topic in the construction of contemporary MMMs given that governments have sometimes abused their powers for identifying misinformation to pursue agendas designed to counter any resistance and criticism. While there are no explicit applications of MMMs in government usage that I am aware of, the potential for their application alone is worthy of discussion given looming controversies in recent misinformation legislation. To illustrate, the Singaporean government had legislated the Protection from Online Falsehoods and Manipulation Act 2019, which intends to "prevent the electronic communication in Singapore of false statements of fact, to suppress support for and counteract the effects of such communication, to safeguard against the use of online accounts for such communication and for information manipulation, to enable measures to be taken to enhance transparency of online political advertisements" (Republic of Singapore, 2021). The law allows the government to fine (up to $500,000 SGD) or incarcerate (up to ten years in prison) any individual who publishes material, especially on the internet or through text messages, that can compromise "the security of Singapore...public safety or public tranquility," could "prevent any influence of the outcome of an election to the office of President" or could "prevent a diminution of public confidence in the performance of any duty or function of, or in the exercise of any power by, the Government." The law states explicitly that this can also apply to those who reside outside of Singapore as well (11).

Notice that this entails that it is potentially possible for researchers creating MMMs to be convicted of spreading 'a false statement of fact' in such a regime given that MMMs' construction sometimes relies upon crowdsourced epistemic judgments as to the truth-value of information. Nowhere in the document is a 'false statement' defined, rendering the law dangerous in its vagueness, especially considering Singapore's history of repressing journalists, ranking 149th out of 179 countries in the 2022 Reporters Without Borders index (Reporters Without Borders, 2022). For example, the law has already been used to pursue legal action against an anti-vaccination website alleging it is publishing false information about the safety of vaccines (Berger, 2021). And yet, what counts as a sufficient threshold for vaccine safety may justifiably be a subjectively grounded preference admitting of a legitimate plurality of reasonable positions. This depends on one's risk assessments and so illustrates how a judgment about the

purported falsity of an information source's recommendation towards vaccine hesitancy is really just a matter of disagreement concerning epistemic risk rather than merely truth. This example therefore illustrates the potentially grave consequences of ignoring sensitivities of the epistemic supply chain in the construction of MMMs should they be used to flag misinformation in digital channels.

5 Conclusion

MMMs have been constructed for nearly two decades and while their mathematical foundations and engineering are well understood, their epistemic workflow remains under analyzed and philosophically inadequate. Given that the annotations of supervised MMMs are typically conducted through a highly heterogeneous epistemic process of consulting the public, journalists, and research experts, the social epistemological consequences of this process render MMMs liable to cognitive biases, abuse, and unclear verdicts as to their construct legitimacy and construct validity. What constitutes misinformation is wholly a matter grounded in procedures that are irreducibly social insofar as citizens' conceptions of what misinformation is ought to be factored into account in the construction of MMMs in more sophisticated ways than are currently practiced. MMMs exhibit disguised features of citizen science which affect the assessment of the construct legitimacy and construct validity of MMMs that is more complex than merely satisfying a set of statistical performance criteria.

Acknowledgements I thank the following for constructive feedback on ideas in this paper: Brian Baigrie, 921 Franz Huber, Michael Miller, Regina Rini, Denis Walsh, the Pittsburgh HPS fringe theory group, the York 922 University moral psychology lab, four anonymous reviewers, the Hong Kong Catastrophic Risk Centre for funding, and the Philosophy of Contemporary and Future Science research group at Lingnan University, Department of Philosophy. All errors and infelicities are mine alone.

References

Abouzeid, A., Granmo, O. C., Webersik, C., & Goodwin, M. (2021). Learning automata-based misinformation mitigation via hawkes processes. *Information Systems Frontiers, 23*, 1169–1188.

Alenezi, M. N., & Alqenaei, Z. M. (2021). Machine learning in detecting COVID-19 misinformation on twitter. *Future Internet, 13*(244), 1–20.

Alexandrova, A. (2017). *A philosophy for the science of well-being*. Oxford University Press.

Arendt, H. [1953] (1976). *The origins of totalitarianism*. Houghton Mifflin Harcourt Publishing Company.

Berger, M. (2021). Singapore invokes 'fake news' law in push against anti-vaccine website. *Washington Post*. https://www.washingtonpost.com/world/2021/10/25/singapore-fake-news-law-anti-vaxxer-coronavirus/. Accessed 4 Sept 2022.

Brennan, J. (2016). *Against democracy*. Princeton University Press.

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting*. 3rd Edition. Springer.

Caled, D., & Silva, M. J. (2022). Digital media and misinformation: An outlook on multidisciplinary strategies against manipulation. *Journal of Computational Social Science, 5*, 123–159.

Castillo, C., Mendoza, M., & Poblete, B. (2012). Predicting information credibility in time-sensitive social media. *Internet Research, 23*(5), 560–588.

Chatham, M. L. (2008). The death of George Washington: An end to the controversy? *The American Surgeon, 74*(8), 770–774.

Cheng, L., Guo, R., Shu, K., & Liu, H. (2021). Causal understanding of fake news dissemination on social media. *KDD '21: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 148-157.

Chronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Coan, T. G., Boussalis, C., Cook, J., & Nanko, M. O. (2021). Computer-assisted classification of contrarian claims about climate change. *Nature Scientific Reports, 11*(22320), 1–12.

de Ridder, J. (2022). Online illusions of understanding. *Social Epistemology*. https://doi.org/10.1080/02691728.2022.2151331

Dretske, F. (1983). Précis of knowledge and the flow of information. *Behavioral and Brain Sciences, 6*(1), 55–90.

Du, J., Preston, S., Sun, H., Shegog, R., Cunningham, R., Boom, J., Savas, L., Amith, M., & Tao, C. (2021). Using machine learning-based approaches for the detection and classification of human papillomavirus vaccine misinformation: Infodemiology study of reddit discussions. *Journal of Medical Internet Research, 23*(98), 1–12.

Elgin, C. (2017) True enough. MIT Press.

Fallis, D., & Mathieson, K. (2019). Fake news is counterfeit news. *Inquiry*, 1-20.

Fallis, D. (2015). What is disinformation? *Library Trends, 63*(3), 401–426.

Feest, U. (2020). Construct validity in psychological tests - the case of implicit social cognition. *European Journal for Philosophy of Science, 10*(4), 1–24.

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

Franklin, A. (2016). *What makes a good experiment?* The University of Pittsburgh Press.

Gillies, D. A. (1971). A falsifying rule for probability statements. *British Journal for Philosophy of Science, 22*, 231–261.

Goldenberg, M. (2021). *Vaccine hesitancy*. University of Pittsburgh Press.

Government of Canada. (2022a). Online disinformation. Government of Canada. https://www.canada.ca/en/canadian-heritage/services/online-disinformation.html. Accessed 31 Aug 2022.

Government of Canada. (2022b). Canada's efforts to counter disinformation - Russian invasion of Ukraine. Government of Canada. https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/response_conflict-reponse_conflits/crisis-crises/ukraine-disinfo-desinfo.aspx?lang=eng

Gruppi, M., Horne, B. D., & Adali, S. (2021). Workshop proceedings of the 15th international AAAI conference on web and social media. *Association for the Advancement of Artificial Intelligence*, 1-10.

Guerrero, A. (2014). Against elections: The lottocratic alternative. *Philosophy & Public Affairs, 42*(2), 135–178.

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances, 5*(5686), 1–8.

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

Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., & Bonn, A. (2018). *Citizen science: Innovation in open science, society and policy*. UCL Press.

Horne, B. D. (2020). *Robust news veracity detection*. Rensselaer Polytechnic Institute. Dissertation.

Horne, B. D., Gruppi, M., & Adali, S. (2020). 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*, 1-4.

Hou, R., Pérez-Rosas, V., Loeb, S., & Mihalcea, R. (2019). Towards automatic detection of misinformation in online medical videos. *Proceedings of the 2019 International Conference on Multimodal Interaction*, 235–243.

Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: A survey and new perspectives. *Social Network Analysis and Mining, 10*(82), 1–20.

Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2017). Novel novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia, 19*(3), 598–608.

John, O. P., & Soto, C. J. (2007). The importance of being valid: Reliability and the process of construct validation. In Richard W. Robins, R. Chris Fraley, & Robert F. Krueger (Ed.), *Handbook of research methods in personality psychology*, (pp. 461–94). Guilford.

Khan, J. Y., Kohndaker, M. T. I., Afroz, S., Uddin, G. & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications, 4*(100032), 1–12.

Laudan, L. (1981). A confutation of convergent realim. *Philosophy of Science, 48*(1), 19–49.

Longino, H. (2022). What's social about social epistemology? *Journal of Philosophy, 119*(4), 169–195.

Mahase, E. (2021). Covid-19: US suspends Johnson and Johnson vaccine rollout over blood clots. *British Medical Journal, 373*(970), 1.
Mishra, S., Shukla, P., & Agarwal, R. (2022). Analyzing machine learning enabled fake news detection techniques for diversified datasets. *Wireless Communications and Mobile Computing,1–18*. https://doi.org/10.1155/2022/1575365
Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. The MIT Press.
Nevo, D., & Horne, B. D. (2022). How topic novelty impacts the effectiveness of news veracity interventions. *Communications of the ACM, 65*(2), 68–75.
Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme, 17*(2), 141–161.
Oreskes, N. (2004). The scientific consensus on climate change. *Science, 306*(5702), 1686.
Osman, M., Adams, Z., & Meder, B. (2022). People's understanding of the concept of misinformation. *Journal of Risk Research, 25*(10), 1239–1258.
Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences, 25*(5), 388–402.
Reporters Without Borders. (2022). The ranking. *Reporters Without Borders*, https://rsf.org/en/ranking. Accessed 4 Sept 2022.
Republic of Singapore. (2021). Protection from online falsehoods and manipulation act 2019. *The Statutes of the Republic of Singapore*.
Robinson, L. D., Cawthray, J. L., West, S. E., Bonn, A., & Ansine, J. (2018). Ten principles of citizen science. In S. Hecker, M. Haklay, A. Bowser, Z. Makuch, J. Vogel, & A. Bonn (Eds.), *Citizen science: Innovation in open science, society and policy* (pp. 1–3). UCL Press.
Shao, C., Hui, P. M., Wang, L., Jiang,X., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2018). Anatomy of an online misinformation network. *PLoS ONE, 13*(4), 1–23.
Shu, K., & Liu, H. (2019). *Detecting fake news on social media*. Morgan & Claypool.
Søe, S. O. (2018). Algorithmic detection of misinformation and disinformation: Gricean perspectives. *Journal of Documentation, 74*(2), 309–332.
Sorokin, A., & Forsyth, D. (2008). Utility data annotation with amazon mechanical turk. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 1-8. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4562953. Accessed 11 Oct 2022.
Stone, C. (2019). A defense and definition of construct validity in psychology. *Philosophy of Science, 86*, 1250–1261.
Swire-Thompson, B., & Lazer D. (2020). Public health and online misinformation: Challenges and recommendations. *Annual Review of Public Health, 41*, 433–451.
Tromble, R. (2021). Where have all the data gone? A critical reflection on academic digital research in the post-API age. *Social Media + Society*, *7*(1), 1–8.
United Nations Human Rights Council. (2018). Report of the detailed findings of the independent international fact-finding mission on Myanmar. *United Nations Human Rights Council*. https://www.ohchr.org/sites/default/files/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_CRP.2.pdf, Accessed 3 Oct 2022.
van Fraassen, B. (1980). *The scientific image*. Oxford University Press.
Vosoughi, S., Ry, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*, 1146–1151.
Yee, A. K. (2023). Information deprivation and democratic engagement. *Philosophy of Science*, *90*(5).
Zubiaga, A., Liakata, M., Proctor, R., Wong Sak Hoi, G., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE, 11*(3), 1–29.