Ethics and Information Technology (2024) 26:2
https://doi.org/10.1007/s10676-023-09734-6

ORIGINAL PAPER

Can machine learning make naturalism about health truly naturalistic? A reflection on a data-driven concept of health

Ariel Guersenzvaig[1]

Accepted: 7 November 2023 / Published online: 12 December 2023

Abstract
Through hypothetical scenarios, this paper analyses whether machine learning (ML) could resolve one of the main shortcomings present in Christopher Boorse's Biostatistical Theory of health (BST). In doing so, it foregrounds the boundaries and challenges of employing ML in formulating a naturalist (i.e., *prima facie* value-free) definition of health. The paper argues that a sweeping dataist approach cannot fully make the BST truly naturalistic, as prior theories and values persist. It also points out that supervised learning introduces circularity, rendering it incompatible with a naturalistic perspective. Additionally, it underscores the need for pre-existing auxiliary theories to assess results from unsupervised learning. It emphasizes the importance of understanding the epistemological entanglements between data and data processing methods to manage expectations about what data patterns can predict. In conclusion, the paper argues against delegating the final authority for defining complex concepts like health to AI systems, as it necessitates ethical judgment and capacities for deliberation that AI currently lacks. It also warns against granting creators and deployers of AI systems the discretionary authority to determine these definitions outside the wider social discussion, advocating for ongoing public engagement on normative notions. Failure to do so risks limiting individuals and collectives' ability to shape a just digital future and diminishes their fundamental epistemic agency.

Keywords Machine learning · Health · Theory · Normative concepts · Big data · Epistemology

Introduction

This paper critically examines the hypothetical application of big data and machine learning for epistemic purposes. It assesses its philosophical entailments through an exploration of conceptual scenarios that enable forging theoretical connections between two separate domains—the philosophy of the biomedical sciences and the philosophy of data and artificial intelligence.

These scenarios speculate on if machine learning (ML) could be employed to resolve some of the shortcomings present in the Biostatistical Theory of Health (BST), a vigorously debated naturalist theory of health proposed by Boorse (1977, 1997, 2014). The BST seeks to be value-free and shielded from prior theoretical assumptions, solely relying on empirical facts to define health—which is conceptualized by Boorse as statistically normal functioning. For its many critics, however, it fails to be naturalist in the strict sense because prior norms and theories inevitably creep in. While this is the topic I wish to address, let me be clear at the outset that my goal here is not to ascertain whether ML can *actually* provide a solution to BST's problems. Instead, the conceptual scenarios—which are not fully-blown thought experiments—serve as a sort of *leitmotif* to analyze, clarify, and explore various philosophical questions raised by the use of ML for epistemic purposes.

This exploration is set against the backdrop of *dataism*, a view that holds that data 'is a transparent and reliable lens that allows us to filter out emotionalism and ideology; that data will help us […] foretell the future' (Brooks, 2013). More specifically, it is set within the context of the 'end of theory' perspective, a radical dataist view holding that data processing and correlation analysis can lead directly to knowledge without the need for prior theories, not only in the natural sciences but also in social investigations (Anderson, 2008). This data-driven approach seems to be, *prima*

✉ Ariel Guersenzvaig
aguersenzvaig@elisava.net

[1] Elisava Barcelona School of Design and Engineering, University of Vic-UCC, Barcelona, Spain

facie, methodologically adequate to resolve some of the shortcomings present in BST.

The thrust of my argument is that not even a sweeping *dataist* approach could rescue Boorse's biostatistical theory of health and make it truly naturalist because prior theories and values are ineradicable *in principle*—at least in the scenarios I develop. In other words, Boorse's theory of health cannot be shielded from values by using inherently value-laden technologies that mediate and co-produce our descriptions and interpretations of the world.

What, a reader may ask, is the relevance of all this if the argument about the non-normativity of data is well-known, especially among the readers of this journal?[1] My reply would be that this exploration could be certainly of interest to those who hold a naive view about dataism, but also to those who are more familiar with this criticism. By analyzing and foregrounding the boundaries and challenges of employing ML in the formulation of a naturalist definition of health, this paper highlights the normativist nature of ML from a practical and context-specific perspective. The goal here is not to establish that data is intrinsically normativist; indeed, many readers already know this. Instead, it is to provide a novel case in point that clarifies and elaborates on how normativism comes into play even when employing an assumed naturalism approach.

This is all too relevant because despite well-known retorts against the putative neutrality of data, statistics is evolving into data science and artificial intelligence (AI) systems are becoming increasingly ubiquitous in high-stakes contexts. The dataist view is 'ruling the world', as Porter (2020, p. XIV) argues: '[s]tatistical routines have been put to work in therapeutic medicine, studies of classroom effectiveness, policing, development economics, and ten thousand rankings and metrics. All have been seriously criticized, yet they continue to be marketed, indeed ubiquitously, as the holy of holies'. Also, amid concerns about reproducibility and overoptimism (Kapoor & Narayanan, 2022), ML algorithms commence to be used for the production of scientific knowledge. Consider the use of 'digital simulacra' in biomedicine, which indicate a potential abandonment of the concepts of causality and representation in favor of the epistemology of data-first approaches and predictive modelling using ML (Cho & Martinez-Martin, 2022).

Along these lines, this inquiry is relevant because concepts such as health are functionally normative—*they guide how we live*. To say that something is 'healthy' activates a plurality of meanings and connotations according to socially embedded guidelines and traditions concerning what is to be preferred and what is to be avoided. The powerful rhetorical mechanism of dataism conceals the prior normative presence of some picture or another of what it means to be healthy. This concealment occurs through a narrative of rigor, value-neutrality, and numerical objectivity that forgoes the true methodological and epistemological complexities of ML. Moreover, it is important to make explicit and re-emphasize the epistemological entanglements between data and data processing methods in order to temper expectations regarding what can be predicted and explained on the basis of data patterns, as well as to cast doubt on the contribution that ML and data-first approaches can make to naturalism about health.

In sum, it is not my intention to rehearse an abstract critique of dataism. Rather, I will take its strongest version—the end of theory view—at face value, engaging and questioning it obliquely by examining the conceptual integrity and entailments of its application. Neither is my intention to present a general critique of the numerous challenges in the application of ML to social issues. ML systems have been repeatedly shown to be plagued by serious socio-ethical problems and there is a vast quantity of sources that deal with these issues, which I will discuss only in so far as they are connected to the issue of normativism *vis a vis* naturalism.

The content proceed as follows: Sect. "Dataism: from statistics to end of theory" introduces the end of theory perspective, especially for those unfamiliar with the philosophy of (big) data and sketches two types of dataism. Section "Boorse's concept of health and telling who's healthy" summarizes the biostatistical theory of health and a main shortcoming. The scenarios are introduced and discussed in Sect. "Machine learning to the rescue?". Section "Conclusion: correlations are not enough for a theory of health" offers conclusions.

Dataism: from statistics to end of theory

The origins of dataism can be traced back to the early 19th century, when the likes of Quetelet[2] and Galton[3], following the inductive reasoning promoted by philosophers like Bacon, and drawing on the empirical successes of 17th century natural scientists, executed a true epistemic revolution by thinking of statistical patterns as inherently explanatory (Hacking, 1990; Porter, 2020). These thinkers established

[1] I thank one of the anonymous reviewers for raising this concern.

[2] Adolphe Quetelet (1796–1874) was a Belgian astronomer, mathematician, and sociologist responsible for introducing statistical methods to the social sciences.

[3] Francis Galton (1822–1911) was a British statistician, sociologist, and polymath who is famous for being the father of eugenics and scientific racism, as well as for describing the statistical notions of correlation and regression, and the phenomenon of regression towards the mean.

quantitative and statistical reasoning as a legitimate mode of social inquiry while also seeking to render theoretical understanding unnecessary. Moreover, they sought to eradicate the very notion of causation—which was too metaphysical and therefore, in their eyes, unscientific[4]—to replace it with *laws* of human nature based on correlations found in empirical data in the spirit of Newtonian mechanics. The more data about the world, the more inductions and generalizations one can make and thus the more laws one can establish (Hacking, 1990, pp. 62–63). Mathematized science strengthened this by taking statistical methods as instruments for both knowledge and proof, something that manifestly surfaces in BST.

During the last two centuries, data collection and statistical analyses affecting all areas of public and private life have enabled theory creation and revision in the social sciences and vertebrated public debate (Desrosières, 1993; Porter, 2020). Data—understood 'as an abstraction from or a measurement of a real-world entity such as persons, objects, or events' (Kelleher & Tierney, 2018 p. 39)—in expressions like 'this is backed by data' became a common way to legitimize claims and decisions about poverty, education, employment, and virtually any other facet of social life. This was only accentuated by the widespread dissemination of computers and databases, and the rise of big data, which is characterized by 'the extreme volume of data, the variety of the data types, and the velocity at which the data must be processed' (Kelleher & Tierney, 2018 p. 9). Today, large data sets are being used for discovery and prediction by businesses and all types of organizations (Mayer-Schonberger et al., 2013; Holmes, 2017).

Two types of dataism

I will categorize dataism into two types: *weak* and *strong*. The difference between the two is rather simple. *Strong* dataism ostensibly claims to be data-driven and hypothesis-free, while *weak* dataism admittedly rests on a theory or a hypothesis that vertebrates the computation. Though our focus will be on strong dataism, weak dataism will be introduced next to clarify the difference.

Common AI-based emotion-recognition systems are a good example of weak dataism. These systems purportedly recognize emotions, affective states, and even intentions by analyzing, for instance, recordings of the interaction of students in a classroom or video interviews of job candidates. These systems exemplify weak dataism because they are clearly underpinned by theoretical models such as FACS (Facial action coding system), a taxonomy of universal emotions (happiness, surprise, anger, disgust, sadness and fear) developed by psychologist Paul Ekman (Ekman & Rosenberg, 2005). Or consider iBorderCtrl, whose goal is to detect 'bona fide' and 'non-bona fide' travelers at border crossings by detecting indicators of deception during interviews (Sánchez-Monedero & Dencik, 2022). I see this as another example of weak dataism because the system is not hypothesis-free as it presupposes that people produce non-verbal facial micro-expressions when they lie. For the makers of iBorderCtrl, detecting these micro-expressions is equivalent to detecting deception.

Surely, there are many strong *prima facie* reasons for not deploying systems like these. They are brittle, ideological, and possibly pseudoscientific and racist (see Crawford, 2021; Feldman Barret et al., 2019). But this is a different issue. The central point is that the theory that supports these systems might be radically flawed, yet it explicitly drives the necessary statistical analysis.

Strong dataism

Chris Anderson's essay 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete' (Anderson, 2008)[5]—a highly influential argument cited more than three thousand times according to Google Scholar—epitomizes strong dataism. Anderson argues that using big data and artificial intelligence (AI) for epistemic purposes will make *ex-ante* theories and hypotheses redundant. He doesn't just propose to use AI to computationally support scientific discovery and theory generation but wants it to take the lead because 'science can advance even without coherent models'. Data-driven discovery can replace theory in both the natural and the social sciences. In his own words:

> This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.

In strong dataism, correlations are enough. Anderson posits that if, for instance, Google says that this page is

[4] This empiricist view persisted until the emergence of the Vienna Circle in the 1930s, when the neopositivists had to deal with things and phenomena that were not conclusively verifiable by confrontation with direct experience such as bacteria and quantum mechanics. Eventually, the requirements became more lenient and a tripartite classification of all our putative judgments was introduced to accommodate more than. For more historical background and a masterful treatment of this question, see Putnam, (2002).

[5] All references to Anderson from this point onwards refer to (Anderson, 2008).

better than that one, as long as 'the statistics of incoming links say it is, that's good enough. No semantic or causal analysis is required'. We don't have to know *why*. We don't even need a hypothesis to control empirically. The gist of the end of theory is that we just need to know *that* something is the case. Anderson subverts thus the general tenets of mid-20th century—i.e., Popperian—science, in which the itinerary of discovery starts by formulating tentative hypothesis (driven by conjectures grounded on previous theory and data) that are validated empirically, yielding thus new data that serves to accept, amend or reject the hypothesis (Popper, 2002).

In line with Anderson, authors such as Mayer-Schonberger and Cukier (2013) also argue in favor of focusing on correlations and less on causality. Steadman (2013) also rehearses Anderson's claims: '[a]lgorithms will spot patterns and generate theories, so there's a decreasing need to worry about inventing a hypothesis first and then testing it with a sample of data'.

More recently, data-driven discovery has been defended by the astrophysicist Kevin Schawinski: 'Let's erase everything we know about astrophysics. To what degree could we rediscover that knowledge, just using the data itself?' (cited in Falk, 2019). Elsewhere, Schawinski et al. (2018) propose a 'generative' approach that presents a more plausible version of Anderson's argument by conceding that human insight is still required for high-level interpretation, which enables an expert to make sense of the discoveries. For some, an instantiation of this perspective can be found in the case of *AlphaFold*, an AI system that has been able to accurately predict the 3D structure of a protein, thus solving one of the great contemporary challenges of biology (Heaven, 2020).

Several authors have engaged with the implications of the 'data deluge' for science and epistemology. Kitchin (2014, p. 5) argues that 'whilst data can be interpreted free of context and domain-specific expertise, such an epistemological interpretation is likely to be anaemic or unhelpful as it lacks embedding in wider debates and knowledge'. boyd & Crawfordl (2012) critique Anderson's 'sweeping dismissal of all other theories and disciplines [as] it reveals an arrogant undercurrent in many Big Data debates where other forms of analysis are too easily sidelined'. These authors question big data's assumptions and biases, dismissing a purely inductive science. They also spurn the sharp dichotomy between the theoretical and the empirical that Anderson suggests as it hampers a correct understanding of the constructivist dimension of empirical data. Naturally, the standard critique against quantitative reasoning and its alleged rigor, value-neutrality, and objectivity can be also marshaled against Anderson's ideas (see Hacking, 1990; Desrosières, 1993; Porter, 2020).

Strong dataism as an emergent reality

While we can safely presume that not all AI systems are developed and marketed by strong dataists who believe that correlations are enough, there are, at least as I see it, sufficient real cases to assert that the end-of-theory is not a fringe perspective but one that merits a reprise of the critic against it. If only as a reminder to vehemently question many of the claims of objectivity, value neutrality, rigor, and scientific fairness often made by vendors of data-driven systems and even by academic researchers. I will summarily introduce two controversial cases to illustrate this point.

The first example: two Stanford researchers (Wang & Kosinski, 2018) used machine learning (deep neural networks performing logistic regression) to analyze more than 35,000 facial images to classify these according to sexual orientation; i.e., to distinguish between gay and heterosexual persons. To achieve this purpose the classifier included fixed (e.g., nose shape) and transient facial features (e.g., grooming style). The authors contend that 'faces contain much more information about sexual orientation than can be perceived and interpreted by the human brain'. Elsewhere, Kosinski, one of the authors, claims that the 'whole idea of machine learning is that you can train it on the original sample, and then as the machine works, it will just start matching patterns and noticing patterns and enriching the model' (cited in Resnick, 2018).

The second example is an AI system that, according to its creators (Wu & Zhang, 2016), can predict if a person is a convicted criminal with nearly 90% accuracy by using facial analysis The system was trained using ID-style face photos of people previously convicted of violent and non-violent crimes, and pictures from people without convictions harvested from the internet. This data set was divided into two subsets: a positive and a negative class, for criminals and non-criminals, respectively. One of the conclusions is that 'the faces of general law-biding public have a greater degree of resemblance compared with the faces of criminals, or criminals have a higher degree of dissimilarity in facial appearance than non-criminals.' While the study wasn't peer-reviewed but published in the popular pre-print repository arXiv, it received massive attention (for a demolishing rebuttal see Agüera y Arcas et al., 2018). In characteristically dataist fashion, Wu & Zhang (2016, p.2) praise machine learning systems because:

> Unlike a human examiner/judge, a computer vision algorithm or classifier has absolutely no subjective baggage, having no emotions, no biases whatsoever due to past experience, race, religion, political doc-

> trine, gender, age, etc., no mental fatigue, no preconditioning of a bad sleep or meal.

Wu & Zhang (2016, p. 6) ask themselves 'what features of a human face betray its owner's propensity for crimes?', and the answer they provide paradigmatically illustrates the end-of-theory mindset and its methodological naturalism:

> We try to answer the question in the most mechanical and scientific way allowed by the available tools and data. The approach is to let a machine learning method explore the data and reveal the most discriminating facial features that tell apart criminals and non-criminals.

Relatedly, in May 2020, two professors and a graduate student claimed that their algorithm could infer criminality from faces. The claims were made in a paper accepted at a peer-reviewed conference for which the major academic publisher Springer Nature would publish the proceedings. After a letter from scholars from fields such as computer science, ethics, and sociology condemning the paper's 'unsound scientific premises, research, and methods' was released, Springer confirmed it would not publish it (Hatmaker, 2020).

Naturally, all this is deeply problematic, and there are strong reasons not to develop systems like these (see Agüera y Arcas et al., 2018; Pasquale, 2018). But the point I want to underscore with these examples is that despite being scientifically and ethically problematic to reason from the unsupported and defective assumption that the numbers speak for themselves, this still occurs. It remains therefore strategic to once again question the end of theory view, which is, alas, not circumscribed to a provocative piece printed in a popular magazine but can already be found in science and embedded in commercial products.

In Sect. "Machine learning to the rescue?", aided by hypothetical scenarios, I will have more to say about the shortcomings of strong dataism. But first, I turn to Boorse's concept of health and briefly introduce the ideas that will be discussed in the scenarios.

Boorse's concept of health and telling who's healthy

The philosophy of health is a vast topic and there is no definite consensus on what health is (for a comprehensive point of departure see Murphy, 2023). In the Western literature, we find, on the one hand, normativist theories (also known as constructivist), whereby health is essentially value-laden. To say something is healthy is not only to make a description of a biological state but also to make an evaluation of it (i.e., it is good to be healthy and bad to be diseased). On the other hand, we encounter naturalist theories (also known as non-normativist) such as Boorse's biostatistical theory, which, to reiterate, will be our focus. In naturalist theories, health is a value-free theoretical notion determined by empirical facts. Naturalist accounts aim to give an objective definition of what health and disease are without involving evaluative judgments grounded on prior theoretical assumptions, in a way similar to the type of definitions we encounter in the natural sciences. As will become clear shortly, the BST tones with Anderson's strong dataism in that it seeks to derive knowledge solely from statistical data. Now we turn to BST's main elements and one major objection posed against it, which will be the focal point of the discussion during the hypothetical scenarios.

The BST in a nutshell

The biostatistical theory rests on a non-normative understanding of biological function and a statistical notion of 'normality'. Boorse (1977, p. 542) posits that health is the 'statistical normality of function' and that 'the normal is the natural' (1977, p. 554). For Boorse, health and disease are just biological states: 'if diseases are deviations from the species biological design, their recognition is a matter of natural science, not evaluative decision' (Boorse, 1977, p. 543). To say that an organism is healthy is to describe a natural fact, not to make an assessment of it in terms of good or bad, desirable or undesirable, and so on.

Diseases are 'internal states that depress a functional ability below species-typical level' (1977, p. 542; 2014, p. 684). An organism is thus healthy when its functioning conforms to its natural design and function. Health is the fitness of an organism to perform its normal functions with statistically normal efficiency under typical conditions. For Boorse (1977, p. 555) normal function 'is a statistically typical contribution by it to their individual survival and reproduction'. Typical contributions are those 'within or above some chosen central region of their population distribution' (Boorse, 1977, pp. 558–559), i.e., those close to the statistical mean.

Although 'normal' levels could be determined statistically for the whole species, a comparison at a species level would be clinically inoperative because a species' functional design seems to be contingent on sex, age, and race (Boorse, 1977, p. 558). Hence the statistical abstractions should be made from classes smaller than species. The upshot is that to assess the normality of a biological state for a subgroup within a species, Boorse needs some sort of benchmark of normality against which things are compared. To this end, he introduces the notion of 'reference class', which is a subset of the whole species. A reference

class is 'a natural class of organisms of uniform functional design; specifically, an age group of a sex of a species' (1977, p. 555). For example, 'a 35 years old white woman' or 'a male neonate of Aymara ancestry'. For Boorse, if we want to establish the health of a neonate's heart, we should compare it to other neonates, factoring in 'sex and race'[6] as well (1977, p. 558), and not to an average adult human heart as an adult with the constant heart rate of a neonate would be considered diseased, and vice versa. Boorse's theory is much richer than I can cover here, but this is the rough idea.

Kingma's objection

Over the years, many authors have criticized the BST (e.g., Ereshefsky, 2009; Gammelgaard, 2000; Kingma, 2007; Law & Widdows, 2008) and Boorse has produced extensive rebuttals (Boorse, 1997, 2014). Next, I will introduce the essence of one specific objection, which will concentrate our attention during the scenarios.

However reasonable and clinically relevant reference classes may be, Boorse undermines himself methodologically—and rather evidently so—by introducing *certain* reference classes, which necessarily presuppose a prior conception of health. This is the thrust of an important objection noted by Kingma (2007, 2014). For Kingma it is unclear *why* it would be appropriate from a naturalist perspective to factor sex, age, and race instead of other criteria to calculate normality. In other words, there are no empirical facts that sufficiently determine that 'male white neonates' is an appropriate reference class, but 'heavy smokers' or 'children with dental cavities' are not.

Smoking and caries are statistically common, yet we would reject adopting 'children with dental cavities' or 'heavy smokers' as a reference class because these classes would contradict our most basic intuitions. What's more, cancer, heart disease, and lung diseases would become 'normal' states when using 'heavy smokers' as a reference class. We do not consider this class appropriate, Kingma argues, because its expected 'normal' functions are clear indicators of disease. This rejection, however, *is* a normative choice that reflects our cultural, political, social, aesthetical, and even, perhaps, religious values.

Kingma (2007, p. 128) warns that 'what it is to be healthy is not to be normal with respect to *any*[7] reference class, but to be normal with respect to "appropriate" reference classes only'. When the right reference classes are used for an evaluation Boorse's theory offers an accurate—and even useful—account of health. Nevertheless, Kingma convincingly shows that Boorse cannot justify his choice of appropriate reference classes without involving value judgments and prior theories and conceptions of health and disease. The upshot is that if Boorse's theory seeks to be free from normative knowledge, it should be able to offer a value-free explanation of which criteria constitute an appropriate reference class. It is not enough to *assert* that 'sex', 'race', and 'age' are (the) appropriate reference classes (Kingma, 2007, p. 131). For the biostatistical theory to be truly naturalist and non-circular, the required reference classes must be determined and justified neutrally and empirically objectively without underlying value judgments and intuitions about what it is to be healthy. And this is what the BST fails to achieve.

The problem, I may posit to compound Kingma's objection, is not only in the selection, but also in the value-laden nature of the very criteria proposed by Boorse. Consider 'race', which may have appeared to be an objective notion—a natural kind—during the Enlightenment and particularly from the late 19th century—with Galton—up to the first half of the 20th century, when racial 'science' began to decay after the horrors of Nazism. While it has shaped where we are today and its legacy is still deeply entrenched—scientific racism is a recalcitrant evil—the notion of race has long been discredited as an objective, natural category, and its historical ideological underpinnings have been made explicit (see e.g., Rutherford, 2002; Saini, 2019; Gould, 1996; Lewontin, 1993). Similarly, even allowing for sex to be partially constituted by some empirical indicators such as testosterone levels, its status as a fully-blown natural category has been a hotly debated issue since the 1970s.

Machine learning to the rescue?

If machine learning techniques have been used in attempt to determine what a criminal face looks like and to classify people based on sexual orientation, would it be too much of a stretch to analogously *imagine* that these techniques could also be used to distinguish healthy from unhealthy people? For instance—and I'm paraphrasing Wu & Zhang. (2016, p. 6) cited above—by letting a machine learning method explore the data and reveal the most discriminating features that differentiate healthy and unhealthy people. If this were possible, defining a concept of health around those features would appear closer and this would release the BST from some of the insidious values and prior theories that sabotage its quest toward non-normativity. If we could reasonably hypothesize how the characteristic features of healthy persons could be discovered in the data without prior theory, Boorse's naturalist theory will appear to have one fewer

[6] Boorse's own nomenclature.

[7] Italics in the original.

problem at least until empirical evidence from actual ML systems becomes available.

Next, I will present some scenarios that focus on supervised and unsupervised learning—the two common primary categories of machine learning[8]. These scenarios do not aim at functioning as solid proofs of concept of this application—they are too idealized and simplified for that—but to facilitate an exploration of various philosophical questions related to normativism and naturalism raised by the use of ML.

Scenarios in supervised learning discussed

In supervised learning, the programmers of a system train it by defining a set of expected output results for a range of input data, which are prepared (i.e., labeled) by an individual or a team. Once trained, a model can assign an output label to a new value. The model can be further trained through feedback on whether the assigned label is correct.

For our purposes, a ML system could use any feature of the human body that can be incorporated into database tables: eye color, bone density, hair thickness, lung capacity, skull volumetric measurements, abdominal circumference, blood pressure, and so on. These, in turn, could be used as reference classes for statistical normality. But it would be rather naïve to expect this gambit to solve Kingma's objection, 'Why *these* signs and not others?', we might ask again. Instead of having three attributes or features lacking non-normative justification (sex, race, and age), now we have many more. Prior normativity and subjectivity would still slip through the cracks.

Perhaps there might be a way out of this tangle if we use multiple data sources and 'random sampling' to select the criteria to be used as reference class. The system could be trained with only a subset of the available data, which is randomly selected. Thus, the system may include data about lymph node swelling (or lack thereof), but leave out data about blood pressure, age, height, and so on. This could perhaps bring us closer to assessing an individual's health in a non-normative way. By integrating randomness, Kingma's objection appears to be defused because the selection is based on chance rather than subjective choice or prior theory.

Alas, prior normative influence would not be removed even with a random selection of attributes. The very choice of the data sources to train the system with is based on prior judgments and presumptions. What's more, the pool of raw data the system can choose from determines and restricts *beforehand* the space of potential reference classes that can emerge. None of this render prior theory unnecessary nor bring us any closer to a value-free account of health. To suit this context, we might revise Kingma's objection to 'why *these data sources* and not others?'

A defender of the end-of-theory position might still retort that the system could be trained, at least in theory, using *all* the data that is available about individuals (both medical and non-medical data). In fact, this is precisely what developers of digital simulacra such as the so-called digital twins aim to do. Writing about data minimalization, Cho and Martinez-Martin (2022, p. 49) state that digital simulacra developers 'attempt to collect as much data as possible' precisely to 'avoid making *a priori* assumptions about what data are relevant'.

Yet, even if we use all the data that one can possibly collect, we would still need to label it and this highlights a subjacent and irresolvable problem when integrating supervised learning: *circularity*. Just like a bird-identification app is trained with labels indicating whether a photo contains a type of bird (toucan, flamingo and so on) or a non-bird, the raw data in our system needs to be associated to health through labels so that it can generate an output based on the detected data patterns. If functional, our system could be able to recognize whether a particular individual is healthy or not, and based on this, it could generate reference classes. Yet the notion of health used for labeling would *precede* the creation of the reference class, and this prior distinction between health and disease becomes an insurmountable obstacle for the strong dataist who wants to use supervised ML to improve on the biostatistical theory. Kingma criticizes the BST for being circular, and we see the same occurring when using supervised ML. In short, supervised machine learning could be used to make assessments of health—possibly even valuable ones—but it far from succeeds in making prior theory redundant.

Unsupervised learning scenario discussed

The main idea in unsupervised learning is that the system autonomously detects regularities (i.e., patterns) and recognizes associations between instances in the data (i.e., statistical correlations) without relying on labeling up front.

A strong dataist might therefore contend that unsupervised ML would not be tainted by circularity because the reference classes could simply emerge as theory-free classifications (i.e., clusters) from the data alone, in a unfettered data-driven fashion. Anderson makes this point ardently:

> Petabytes allow us to say: 'Correlation is enough.' We can stop looking for models. We can analyze the data

[8] There are other subfields like reinforcement learning, which I will leave aside as my findings could be extrapolated to it. This also applies to the so-called Large Language Models, which combine unsupervised and supervised learning (fine-tuning), and whose output is generated based on patterns and information present in the text data they were trained on.

> without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

Indeed, enterprises use a classification technique called 'profiling' to create customer segments by detecting patterns present in the data that have not been previously hypothesized. So, profiling could be used to obtain classifications (i.e., reference classes) without the need for causal models. Moreover, as the system would deal with vast purely numerical vectors, whereby the original attributes in the database (i.e., the column labels) wouldn't even be necessary, we could generate reference classes purely with numbers and algorithms.

> But would these classifications truly precede theory? Would they be naturalist *all the way*?

To answer, the discussion needs to examine the nature of data. Given the issue is vast, I will be content to raise reasonable doubts about the possibility of unsupervised ML being able to generate naturalist, atheoretical reference classes. The first things to consider is that data is not directly incorporated into systems as if it was a neutral mirror of empirical reality. This might be well-known to data scholars keen on epistemology and ethics but it needs to be stated and emphasized once again for a wider public.

The data lifecycle is marked by an initial narrowing down of the discovery space guided by prior assumptions. As Kuhn (1977, p. 119) writes: 'To discover quantitative regularity one must normally know what regularity one is seeking and one's instruments must be designed accordingly'. Data must be collected, processed, and made computationally readable, which already implies a transformation of the complexity of the world into database fields. Digital health scholars have highlighted different epistemic styles between the medical and computer engineering communities, even speaking of an outright 'clash of cultures' (Wongvibulsin & Zeger, 2020). Engineers prioritize model performance and may pay less attention to model assumptions and causality, while medical researchers place a higher value on the theoretical state of the art and causal clinical reasoning. The reduction of complexity is thus guided by values such as efficiency, security, autonomy, privacy, cost-effectiveness, and so on. Data is not a Platonic entity, but a construct that is *purposedly made appropriate* to the systems and classification schemes in which they are incorporated. No researcher simply 'throws numbers' into a computer system as Anderson proposes.

Bowker and Star (2000) famously show how classification systems shape and are themselves shaped by worldviews and by social interactions. Categories are never merely naturalist reflections of reality as they expose some aspects and obscure others. Consider the developments in the measuring ideal or abstract properties of a population and how the statistical mean became reified during the 18th and 19th century. Data and presuppositions coalesce around a rhetorical question posed by Hacking (1990, p. 109): 'Why *should*[9] one collect such information [about male height]? It is interesting only if one believes, with Quetelet, that it signifies some underlying real characteristic of a population'

A second problem is sampling bias, which is a common type of data bias among many (Kundi et al., 2022). To exemplify, consider some promising studies of AI systems aim at detecting skin cancer (Esteva et al., 2017; Fink et al., 2020). A big concern is they are more accurate with light skin than with dark skin, which likely has to do with sampling bias (skin tones not properly being taken into account) in the datasets employed to train the models (Adamson & Smith, 2018). Sampling bias is a significant technical flaw in statistical analysis.

Yet, bias may be not just a technical issue related to statistics but a broader one related to ethics. The main raw material used by AI—data—may *systemically* be biased by virtue of being connected to overarching societal issues related to injustice, inequality, and discrimination. There is a vast literature discussing how ethnicity, gender, age, educational level, cognitive abilities, and many other vectors of injustice interact with datasets and algorithms (e.g., Benjamin, 2019; Eubanks, 2018).

Consider gender bias, which affects not only medical data but (Western) medicine itself, whose history shows a structural lack of interest in women's health. Let's look at few examples. 8 of the 10 prescription drugs withdrawn from the U.S. market between 1997 and 2001 posed greater health risks for women than for men (USGAO, 2001). Diseases are ignored when they do not affect men, as with endometriosis (Huntington & Gilmour, 2005). Procedures and therapies might have distinct effects on men and women, which can go unnoticed for many years until women are included in controlled trials (Ridker et al., 2005). The effects of medical interventions on menstruation seem to be an afterthought. Period changes, for instance, were commonly reported after Covid-19 vaccination, but were initially not in the list of common side effects compiled by the UK's regulatory agency MHRA (Male, 2021). We could go on but the upshot is that this disregard for women is mirrored in the data, which offer only a partial and biased knowledge of the world.

[9] Italics in the original.

In this case, the problems we encounter are first and foremost a matter of justice,[10] not only a technical issue related to the internal validity of datasets. Because of systemic injustice, it would be irresponsible to take ML-driven reference classes *as is* without further assessment. This assessment necessarily requires auxiliary theories to ensure they are adequate from functional, statistical, and justice perspectives. Using a 'fairness metric' such as 'demographic parity' might be a way to assess whether the outcomes of a model are distributed fairly among different demographic groups, thus not exhibiting unfair discrimination against certain groups. Yet, the dependency on previous theories is highlighted by the realization that many and mutually incompatible notions of fairness exist (Friedler et al., 2021).

Thirdly, and lastly, descriptive models in the social sciences can change the basic coordinates they describe in a '"double hermeneutic effect", in which an interpretation of the world shapes the very interpretations that comprise it' (Blakely, 2020, p. XXV). Consider how '*the* economy' is measured with *prima facie* neutral metrics such as Gross Domestic Product (GDP), inflation, and unemployment rate, while other indicators such as the humanity of labor, the environmental impact of economic activities, or extreme inequalities are relegated. Yet, which indicator to use is a choice motivated by political and moral views, but at the same time it defines what the economy is and 'how to describe and measure it' (ibid, p. 5). The economy thus *becomes* exclusively what is measured by GDP. In the social realm, there is no escape from normativity.

To summarize, several critical challenges have surfaced when applying the end of theory to a naturalist account of health. Prior theory, as well as notional and axiological subjectivity blight the use of machine learning to improve on Boorse's theory. The labeling required for training data in supervised ML systems introduces an element of circularity that renders categorizations defective. Additionally, machine learning systems are prone to be affected by technical and structural biases. Previous theories and values are also necessary to recognize and mitigate these problems. Concurrently, assessing the appropriateness of a reference class determined by unsupervised ML would also require prior auxiliary theories. Even if machine learning is successful at identifying patterns, it can't be determined without external—normative—intervention whether the detected correlations are meaningful or spurious.

[10] I have distributive and procedural justice in mind but a case can be made for developing ML systems that are also restorative. Dealing with this issue is alas outside the scope of this paper.

Conclusion: correlations are not enough for a theory of health

A key aspect of medical expertise is assigning a normative evaluation such as 'healthy' to a biological state. Take the case of osteoporosis. While its diagnosis largely depends on a quantitative assessment of bone mineral density, the clinical significance of osteoporosis lies in the fractures that arise, and their causes are multifactorial. To assess the risk of fracture there is a myriad of methods with different input variables and models, which generate different risk estimations (Kanis et al., 2017). Different conceptions of health enable health professionals to offer justifying reasons in favor or against calling a state or condition 'healthy' or 'diseased', it's not just a matter of assessing data.

As an additional instance, consider the role obesity, assessed using the Body Mass Index (BMI), plays in establishing the risk of developing coronary heart disease. The BMI is a measurement weight/height^2—that categorizes a person from underweight to obese. The higher the BMI the stronger the cardiovascular risk (Khan, 2018). Even if the data integrated in the index are not biased per se, it's worth noticing that neither are atheoretical in the strict sense as their very existence reflects values and attitudes on health and fatness present in society. Moreover, since the BMI is, for some, but certainly not all medical practitioners, a preferred indicator for detecting risk, it normatively guides prevention and treatment. However, there are other approaches to preventing cardiovascular risk such as the 'Social Determinants of Health' (SDH) framework. While this framework does not necessarily reject the value of individual risk indicators such as the BMI, it pays primary attention to the upstream social determinants in which the individual measurements are nested. In other words, instead of dividing a person's weight by the square of their height, the SDH framework focuses on systemic and structural parameters such as access to good transportation, education, and housing, which can also be positively or negatively linked to heart disease and stroke (see e.g., WHO, 2010).

These critical deliberations on whether to call a biological state such as a high BMI a strong risk marker for poor cardiovascular prognosis or to define widely prevalent aspects of human sexuality such as homosexuality a disorder have profound implications on individuals and society. The most noticeable is their influence on the contents of classificatory standards such as the 'International Statistical Classification of Diseases (ICD) and the 'Diagnostic and Statistical Manual of Mental Disorders' (DSM). To exemplify, whereas hysteria was considered since the second millennium BC a physical disease affecting women especially, it was not until the 19th century when it became a widely diagnosed condition. Yet, more than anything else, hysteria was a reflection

of Victorian gender dynamics and oppressive attitudes toward women. Hysterical neurosis was deleted for the DSM in 1980 (Tasca et al., 2012). Although the medical profession formally no longer uses the term 'hysterical', the punitive effects of hysteria once having been an *official* disease linger on and are daily suffered by women all over the world.

Murphy's (2023) assertion that we now say 'that homosexuality was never a disease, and was just diagnosed on moral grounds' underscores that defining notions such as health and disease is always fraught and value-laden. Such definitions are part of an open debate, not only about *how things are* in actuality but also about *how they should be*. Moreover, the notion of statistical normality, which the BST values greatly, is far from being a neutral and disinterested technique. In the words of Hacking (1990, p. 169), 'the benign and sterile-sounding word "normal" has become one of the most powerful ideological tools of the twentieth century'. Judgements about normality in a population are preceded by a definition of the population and the reference class or classes that should be considered. It is for this reason that selecting a reference class is not merely a matter of establishing positive correlations between data points; it is also a question of ethics and politics. When applied to social issues, choosing a reference class is an inherently political avenue, even when used descriptively, because it inevitably becomes normative, as we discussed above in the context of measuring the economy.

Besides, we would not simply expect a ML system to generate reference classes, which is a computationally straightforward task of finding correlations. Rather, what we would expect is that meaningful correlations are found that enable a system to generate reference classes that are just and clinically *adequate*. To accept as valid the reference classes obtained through the processing of large datasets, we would also want to understand how these were obtained. Some sort of explanatory justifications, in terms of reasons *why* the system generated a particular result, would therefore be required (Casacuberta et al., 2022). We're not solely looking for an account of what computations the system did, like the specific parameter values set by the learning algorithm. We would also seek insights into the *reasons* behind the system's results—explanations that justify it. The question that remains is whether a ML systems can offer such explanations (Coeckelbergh, 2019: pp. 121–122; Campolo et al., 2020, p. 1).

That machine learning won't save the BST from its internal conflicts does not imply that all normatively engaged notions of health are equally coherent or comprehensive. It only emphasizes that any substantive engagement with the notion of health, any classification of a state as healthy or diseased, and, for that matter, also any examination of the clinical adequacy of a reference class necessarily requires the faculty of discretion, a form of judgement that 'combines intellectual and moral cognition' (Daston, 2022, p. 38) to tweak 'the universal law or rule to the particular case' (ibid., p.40). The strong dataism of the end of theory perspective stands opposed to this because whenever a 'reasoning process can be made computable, we can be confident that we are dealing with something that has been universalized' (Porter, 2020, p. 86).

Final remarks

Although a fully-blown injunction against the use of machine learning for epistemic purposes is unwarranted, we should avoid relinquishing to these systems the final say in the formulation of fraught, *thick* concepts such as health and healthy—i.e., concepts that are both descriptive and evaluative (Williams, 2006).[11] Firstly, because these systems overlook and obscure much of the richness, relationality, and subtleties of human existence but first and foremost because formulating fraught notions requires genuine *judgment*—understood as 'deliberative thought, ethical commitment and responsible action'—something no current AI system is capable of (Cantwell Smith, 2019, pp. XV, 82). ML systems—coated with the allure of data-driven objectivity—lack the moral aspect that true discretion demands and must, therefore, not substitute for social deliberation and human judgment.

Secondly, we must not delegate to the creators and deployers of AI systems the discretionary authority to determine meanings and opaquely settle public issues outside the wider social discussion. To do so would be to deny the public the possibility of cocreating a just digital future. Instead, we should underscore the importance that open-ended public deliberation has in a democracy. Democratic societies are said to be characteristically marked by ongoing processes of disagreement and contestation, and defined by 'agonistic' conceptions of the good that emphasize the existence of inherent societal conflicts and antagonisms that must be expressed and debated in a way that doesn't lead to violence (Mouffe, 2000).

I want to conclude with a call to action. We must preserve in the public realm the discretionary power to make judgments about what these fraught, normative notions mean. Safeguarding the ability to evaluate the world through individual and public reasoning, as well as defining 'normality' and the inherently contestable and time-bound reference classes to assess it, entails taking custody of the primacy of the social, deliberative dimension of epistemic agency. This is not only indispensable for democratically navigating

[11] While my focus was on health, I believe the thrust of my arguments could be extended to other *thick* concepts such as happiness and justice.

societal disagreements but also to flourish both as humans and as citizens.

Acknowledgements The author wishes to thank David Casacuberta, Sara Pedraz, and Alger Sans who provided valuable comments on draft versions of this paper. He would also want to thank everyone who contributed feedback at the conferences where this article was presented during the drafting process.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Competing interests The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Adamson, A. S., & Smith, A. (2018). Machine learning and health care disparities in dermatology. *JAMA Dermatology*, *154*(11), 1247–1248. https://doi.org/10.1001/jamadermatol.2018.2348.

Agüera, B., Todorov, A., & Mitchell, M. (2018). *Do algorithms reveal sexual orientation or just expose our stereotypes?* Retrieved October 5, 2022 from https://medium.com/@blaisea/d998fafdf477.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine* Retrieved April 12, 2022 from https://www.wired.com/2008/06/pb-theory/.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim Code*. Polity.

Blakely, J. (2020). *We built reality: How social science infiltrated culture, politics, and power*. Oxford University Press.

Boorse, C. (1977). Health as a theoretical concept. *Philosophy of Science*, *44*(4), 542–573. https://doi.org/10.1086/288768.

Boorse, C. (1997). A rebuttal on health. In J. M. Humber & R. F. Almeder (Eds.), *What is Disease?* (pp. 1–134). Humana Press.

Boorse, C. (2014). A second rebuttal on health. *Journal of Medicine and Philosophy*, *39*, 683–724. https://doi.org/10.1093/jmp/jhu035.

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. The MIT Press.

boyd, & Crawford, K. (2012). Critical questions for Big Data. *Information Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878.

Brooks, D. (2013, Feb 4). The Philosophy of Data. *The New York Times* Retrieved October 1, 2022 from https://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html.

Campolo, A., & Crawford, K. (2020). Enchanted determinism: Power without responsibility in Artificial Intelligence. *Engaging Science Technology and Society*, *6*, 1–19. https://doi.org/10.17351/ests2020.277.

Cantwell Smith, B. (2019). *The promise of artificial intelligence*. The MIT Press.

Casacuberta, D., Guersenzvaig, A., & Moyano-Fernández, C. (2022). Justificatory explanations in machine learning: For increased transparency through documenting how key concepts drive and underpin design and engineering decisions. *AI & Society*. https://doi.org/10.1007/s00146-022-01389-z.

Cho, M. K., & Martinez-Martin, N. (2022). Epistemic rights and responsibilities of digital simulacra for biomedicine. *The American Journal of Bioethics*. https://doi.org/10.1080/15265161.2022.2146785

Coeckelbergh, M. (2019). *AI ethics*. The MIT Press.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Daston, L. (2022). *Rules: A short history of what we live by*. Princeton University Press.

Desrosières, A. (1993). *La Politique Des grands nombres: Historie De La Raison statistique*. Éditions La Découverte.

Ekman, P., & Rosenberg, E. L. (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)* (2nd ed.). Oxford University Press.

Ereshefsky, M. (2009). Defining 'Health' and 'Disease'. *Studies in the History and Philosophy of Biology and Biomedical Sciences*, *40*(3), 221–227. https://doi.org/10.1016/j.shpsc.2009.06.005.

Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). Dermatologist-level classification of Skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Falk, D. (2019). How Artificial Intelligence Is Changing Science. *Quanta Magazine*. Retrieved April 12, 2022 from https://www.quantamagazine.org/how-artificial-intelligence-is-changing-science-20190311/.

Feldman Barrett, L., Adolphs, R., Marsella, S., Martinez, A., & Pollak, S. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychol Sci Public Interest*, *20*(1), 1–68. https://doi.org/10.1177/1529100619832930.

Fink, C., Blum, A., Buhl, T., et al. (2020). Diagnostic performance of a deep learning convolutional neural network in the differentiation of combined naevi and melanomas. *Journal of the European Academy of Dermatology and Venereology*, *34*(6), 1355–1361. https://doi.org/10.1111/jdv.16165.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (Im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the Acm*, *64*(4), 136–143. https://doi.org/10.1145/3433949.

Gammelgaard, A. (2000). Evolutionary biology and the concept of disease. *Medicine Health Care and Philosophy, 3*, 109–116. https://doi.org/10.1023/a:1009999502884

Gould, S. (1996). *The mismeasure of Man, revised edition*. WW Norton & Co.

Hacking, I. (1990). *The taming of chance*. Oxford University Press.

Hatmaker, T. (2020). *AI researchers condemn predictive crime software, citing racial bias and flawed methods* Tech Crunch. Retrieved April 12, 2022 from https://techcrunch.com/2020/06/23/ai-crime-prediction-open-letter-springer/.

Heaven, W. D. (2020). DeepMind's protein-folding AI has solved a 50-year-old grand challenge of biology. *MIT Technology Review*. Retrieved April 12, 2022 from https://www.technologyreview.com/2020/11/30/1012712/deepmind-protein-folding-ai-solved-biology-science-drugs-disease/.

Holmes, D. (2017). *Big data: A very short introduction*. Oxford University Press.
Huntington, A., & Gilmour, J. A. (2005). A life shaped by pain: Women and endometriosis. *Journal of Clinical Nursing*, *14*(9), 1124–1132. https://doi.org/10.1111/j.1365-2702.2005.01231.x.
Kanis, J. A., Harvey, N. C., Johansson, H., et al. (2017). Overview of fracture prediction tools. *Journal of Clinical Densitometry: The Official Journal of the International Society for Clinical Densitometry*, *20*(3), 444–450. https://doi.org/10.1016/j.jocd.2017.06.013.
Kapoor, S., & Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science. Preprint retrieved from https://arxiv.org/abs/2207.07048.
Kelleher, J., & Tierney, B. (2018). *Data science*. The MIT Press.
Khan, S. S., Ning, H., Wilkins, J. T., et al. (2018). Association of body mass index with lifetime risk of cardiovascular disease and compression of morbidity. *JAMA Cardiology, 3*(4), 280–287. https://doi.org/10.1001/jamacardio.2018.0022
Kingma, E. (2007). What it is to be healthy? *Analysis*, *67*(2), 128–133. https://doi.org/10.1093/analys/67.2.128.
Kingma, E. (2014). Naturalism about health and Disease: Adding nuance for progress. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, *39*(6), 590–608. https://doi.org/10.1093/jmp/jhu037.
Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12. https://doi.org/10.1177/2053951714528481.
Kuhn, T. (1977). *The essential tension*. The University of Chicago Press.
Kundi, B., El Morr, C., Gorman, R., & Dua, E. (2022). Artificial Intelligence and Bias: A scoping review. In C. E. Morr (Ed.), *AI and society: Tensions and opportunities* (pp. 199–212). CRC Press.
Law, I., & Widdows, H. (2008). Conceptualising Health: Insights from the Capability Approach'. *Health Care Analysis*, *16*(4), 303–314. https://doi.org/10.1007/s10728-007-0070-8.
Lewontin, R. (1993). *Biology as ideology: The doctrine of DNA*. Harper Perennial.
Male, V. (2021). Menstrual changes after covid-19 vaccination. *Bmj*. https://doi.org/10.1136/bmj.n2211
Mayer-Schonberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live. Work and think*. Hachette.
Mouffe, C. (2000). *The democratic paradox*. Verso.
Murphy, D. (2023). *Concepts of Disease and Health*, [online], in: Zalta, E. & Nodelman, U. (Eds.), The Stanford Encyclopedia of Philosophy (Fall 2023 Edition). Retrieved October 12, 2023 from https://plato.stanford.edu/archives/fall2023/entries/health-disease/.
Pasquale, F. (2018). When Machine Learning is facially invalid. *Communications of the ACM*, *61*(9), 25–27. https://doi.org/10.1145/3241367.
Popper, K. (2002). *The logic of Scientific Discovery*. Routledge.
Porter, T. (2020). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.
Putnam, H. (2002). *The collapse of the fact/value dichotomy and other essays* (pp. 1–45). Harvard University Press.
Resnick, B. (2018). *This psychologist's 'gaydar' research makes us uncomfortable. That's the point* Vox. Retrieved April 12, 2022 from https://www.vox.com/science-and-health/2018/1/29/16571684/michal-kosinski-artificial-intelligence-faces.
Ridker, P. M., Cook, N. R., Lee, et al. (2005). A randomized trial of low-dose aspirin in the primary prevention of Cardiovascular Disease in women. *New England Journal of Medicine*, *352*(13), 1293–1304. https://doi.org/10.1056/NEJMoa050613.
Rutherford, A. (2002). *Control: The dark history and troubling present of Eugenics*. Weidenfeld & Nicolson.
Saini, G. (2019). *Superior: The return of race science 4th state*. Beacon Press.
Sánchez-Monedero, J., & Dencik, L. (2022). The politics of deceptive borders: 'biomarkers of deceit' and the case of iBorderCtrl. *Information Communication & Society*, *25*(3), 413–430. https://doi.org/10.1080/1369118X.2020.1792530.
Schawinski, K., Turp, D. M., & Zhang, C. (2018). Exploring galaxy evolution with generative models. *Astronomy & Astrophysics*, *616*, L4. https://doi.org/10.1051/0004-6361/201833800.
Steadman, I. (2013). Big data and the death of the theorist. *Wired Magazine*. Retrieved July 12, 2022 from http://www.wired.co.uk/news/archive/2013-01/25/big-data-end-of-theory.
Tasca, C., Rapetti, M., Carta, M. G., & Fadda, B. (2012). Women and hysteria in the history of mental health. *Clinical Practice and Epidemiology in Mental Health*, *8*, 110–119. https://doi.org/10.2174/1745017901208010110.
USGAO–United States General Accounting Office. (2001). *Drug Safety: Most Drugs withdrawn in recent years had Greater Health risks for women*. US Government Publishing Office.
Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology, 114*(2), 246–257. https://doi.org/10.1037/pspa0000098
WHO-World Health Organization. (2010). *A conceptual framework for action on the social determinants of health*. World Health Organization. Retrieved April 12, 2022 from https://apps.who.int/iris/handle/10665/44489.
Williams, B. (2006). *Ethics and the limits of philosophy*. Routledge.
Wongvibulsin, S., & Zeger, S. L. (2020). Enabling individualised health in learning healthcare systems. *BMJ Evidence-Based Medicine*, *25*(4), 125–129. https://doi.org/10.1136/bmjebm-2019-111190.
Wu, X., & Zhang, X. (2016). Automated inference on criminality using face images. 4038–4052. https://doi.org/10.48550/arxiv.1611.04135