Three Problems with Big Data and Artificial Intelligence in Medicine

Benjamin Chin-Yee* and Ross Upshur†

ABSTRACT The rise of big data and artificial intelligence (AI) in health care has engendered considerable excitement, claiming to improve approaches to diagnosis, prognosis, and treatment. Amidst the enthusiasm, the philosophical assumptions that underlie the big data and AI movement in medicine are rarely examined. This essay outlines three philosophical challenges faced by this movement: (1) the epistemological-ontological problem arising from the theory-ladenness of big data and measurement; (2) the epistemological-logical problem resulting from the inherent limitations of algorithms and attendant issues of reliability and interpretability; and (3) the phenomenological problem concerning the irreducibility of human experience to quantitative data. These philosophical issues demonstrate several important challenges for these technologies that must be considered prior to their integration into clinical care. Our article aims to initiate a critical dialogue on the impact of big data and AI in health care in order to allow for more robust evaluation of these technologies and to aid in the development of approaches to clinical care that better serve clinicians and their patients.

*Department of Medicine, University of Toronto.

†Department of Family and Community Medicine and Dalla Lana School of Public Health, University of Toronto; Associate Director, Lunenfeld Tanenbaum Research Institute, Sinai Health System, Toronto.

Correspondence: Dr. Benjamin Chin-Yee, St. Michael's Hospital, Department of Internal Medicine, 30 Bond Street, Toronto, ON, M5B 1W8, Canada.

Email: benjamin.chinyee@mail.utoronto.ca.

Perspectives in Biology and Medicine, volume 62, number 2 (spring 2019): 237–256.

We live in the Age of Big Data. In medicine, artificial intelligence (AI) and machine learning algorithms, fueled by big data, promise to change how physicians make diagnoses, determine prognoses, and develop new treatments. An exponential rise in articles on these topics is seen in the medical literature. Recent applications range from the use of deep learning neural networks to diagnose diabetic retinopathy and skin cancer from image databases, to the use of various machine learning algorithms for prognostication in cancer and cardiovascular disease (Esteva et al. 2017; Gulshan et al. 2016; Ow and Kuznetsov 2016; Rumsfeld, Joynt, and Maddox 2016). Many factors are driving the adoption of AI in health care, from the rapid expansion of digital imaging and electronic health records to the development of machine learning algorithms that can perform integrative analysis and are adaptive across multiple applications (Naylor 2018).

Endorsements of the benefits of big data and AI feature in prominent medical journals, along with occasional balanced and critical perspectives.[1] Some proponents claim that "much of the diagnostic and monitoring functions performed by physicians today can be offset to computers and algorithms," envisioning a future where "predictive analytics, using all the relevant metrics for an acute event, such as an asthma exacerbation, seizure, heart failure decompensation, severe depression, or autoimmune disease symptoms, could warn individuals and their physicians before these events occur" (Topol, Steinhubl, and Torkamani 2015, 354). Others go as far as to suggest a future state of data-determinism, where algorithms will generate "perfect risk estimates for individuals" that can "predict with perfect accuracy whether an event would occur or not in every individual" (Sniderman, D'Agostino, and Pencina 2015, 25). Some health-care professionals and policymakers see AI as the panacea for issues facing contemporary health care—the solution that will reduce medical uncertainty and misdiagnosis, find novel therapies for cancer, all while combatting rising costs and overutilization of health-care resources (Amato et al. 2013; Bennett and Hauser 2013; Fleming 2018).

Amid this excitement, the philosophical assumptions that underlie the big data and machine learning movement in medicine are rarely questioned. While a literature on ethical issues arising from big data and AI has emerged (Goodman 2015), further engagement with core philosophical problems facing this movement is required. There is an urgent need for critical scholarship to counterbalance the overexuberance that often surrounds these technologies. To this end, we examine three interrelated philosophical issues that currently confront big data and machine learning in medicine. These issues span major branches of philosophy from logic and epistemology to ontology and phenomenology. The

[1]For examples of endorsements of big data and AI in the clinical literature, see Hinton 2018; Naylor 2018; Sniderman, D'Agostino, and Pencina 2015; Stead 2018; Topol, Steinhubl, and Torkamani 2015. For more balanced opinions, see Beam and Kohane 2018; Obermeyer and Emanuel 2016; Obermeyer and Lee 2017; Thornton 2015. More critical perspectives are offered by Cabitza, Rasoini, and Gensini 2017 and Verghese, Shah, and Harrington 2018.

first problem is epistemological-ontological, concerning the theory-ladenness of data and measurement and how epistemic interests shape big data ontologies. The second problem is epistemological-logical, related to the logical limits of algorithms in modelling clinical complexity and resulting issues of reliability and interpretability for clinical decision making. The third problem is phenomenological, surrounding the irreducibility of human experience to quantitative data that prevents integration into current AI technologies. This discussion reveals some of the philosophical commitments of the big data and machine learning movement in medicine, which demand critical examination to prevent them from potentially misguiding medical research and clinical care.

The Epistemological-Ontological Problem

The first philosophical issue faced by the big data and AI movement in medicine arises at the intersection between ontology and epistemology. It concerns the fundamental ontological question "What sorts of 'things' are there in clinical medicine?" and the epistemological question "How do we generate knowledge of them?" Answers to these questions shape our approaches to data selection and acquisition, and also impact how we represent and characterize clinical phenomena. We explore the epistemological-ontological problem below, beginning with a discussion of the theory-ladenness of data and measurement, followed by an exploration of the implications for big data ontologies.

The Theory-Ladenness of Big Data

Critics of big data have recognized the importance of theory-informed data acquisition and interpretation (Coveney, Dougherty, and Highfield 2016). An underlying assumption of the big data movement is that unbiased, theory-free access to data is possible, and that more data will allow learning algorithms to produce more accurate predictions (Sniderman, D'Agostino, and Pencina 2015). Data, however, does not exist in a vacuum; rather, it represents highly selected information based on a priori assumptions and theories. The relationship between theory and data has been a central issue in the philosophy of science for the past century.

Karl Popper (1962) tells an anecdote from when he was lecturing a group of physics students in Vienna in the 1930s. Popper began his lesson with the following instructions: "Take pencil and paper; carefully observe, and write down what you have observed!" (61). His students were perplexed—what did he want them to observe? As Popper demonstrated in this exercise,

> Clearly the instruction, "Observe!" is absurd . . . . Observation is always selective. It needs a chosen object, a definite task, an interest, a point of view, a problem. And its description presupposes a descriptive language, with property

> words; it presupposes similarity and classification, which in their turn presuppose interests, points of view, and problems. (61)

This problem, sometimes referred to as the "theory-ladenness of observation," occupied many leading philosophers of science throughout the 20th century. Through this example, Popper mounted his critique of logical positivism, a dominant philosophical school at the time, which viewed "facts" as independent, value-free "observational-statements" about the world (Carnap 1919). According to the positivist view, only propositions referring to scientifically verifiable facts are meaningful.

Positivism lives on today in medicine's quantitative epistemologies, of which big data and machine learning are the latest incarnation (Chin-Yee and Upshur 2015, 2018; Goldenberg 2006, 2009). This perspective persists despite serious challenges raised by contemporary philosophy. In addition to Popper, other notable critics have demonstrated how observations always occur in light of an existing "conceptual scheme," or shown how empiric data is shaped by the theoretical assumptions of a prevailing "scientific paradigm" (Kuhn 1962; Quine 1980). Notions of theory-ladenness have been extended to help us understand how other extra-empirical factors shape knowledge production, influencing movements in social epistemology (Longino 2002). Some (but not all) of these factors are captured in Popper's statement on the role of "interests, points of view, and problems" in scientific inquiry.

The implications of these ideas have been applied to medicine, particularly in critiques of evidence-based medicine (Chin-Yee 2014; Goldenberg 2006; Upshur 2005). Although some researchers hold more nuanced understandings of data, by and large the positivist view endures unquestioned in the big data movement. Big data's attitude towards causality is a vestige of logical positivism. Big data and machine learning approaches often neglect causal reasoning, preferring to interrogate data without reference to a causal structure. This may be due to the nature of some machine learning algorithms, such as neural networks, which construct "black boxes" that prevent investigators from identifying causal relationships between inputs and outputs (Tu 1996). Critics argue that lack of a causal structure limits interpretation of data, and that a theoretical framework linking causes to effects underwrites sound clinical reasoning and scientific thought (Pearl 2009). Proponents of big data and machine learning retort that the "atheoretical" perspective is in fact a strength of these methods, avoiding preconceptions that stifle scientific progress. Some pundits have gone as far as to declare the "end of theory" in the era of big data (Anderson 2008).

What these proponents fail to recognize, however, is that there cannot be an atheoretical perspective. Theory is not dead in the era of big data because, as Popper reminded us, data cannot do without theory—all inquiry presupposes "interests, points of view, and problems." This lesson is of particular relevance

to the big data movement. What is observed—what counts as data—is shaped by the epistemic interests of particular investigators and research communities. This is reflected in how data is curated into pipelines to train neural networks and machine learning algorithms.

Consider a common disease such as lung cancer. The "interests" of molecular biologists studying this disease may lie in the role of particular genes in tumor progression; from the "point of view" of clinical researchers, the most relevant observations might be specific clinical endpoints (such as overall survival) measured in randomized controlled trials; and the main "problems" for population health researchers might be to identify important social and environmental determinants of lung cancer incidence or drivers of inequalities in lung cancer survival in order to inform public policy.

What is deemed an "observation" or significant outcome is shaped by epistemic context. In principle, this context dependence need not be rigid and limiting, and differing perspectives can be complementary; however, in practice there is often little meaningful engagement and collaboration across big data research disciplines (Livingstone et al. 2015). The wide range of epistemic interests in biomedical research is demonstrated by the proliferation of a multitude "-omics" disciplines—from the "genome" and "epigenome" all the way to the "interactome" and "exposome"—which collectively lack coherent structural organization across levels to allow for meaningful analysis (Khoury and Galea 2016; Livingstone et al. 2015).

Although the big data movement in health care has ambitions to utilize "all the relevant metrics" mined from electronic health records (Sniderman, D'Agostino, and Pencina 2015), in reality this research utilizes narrower definitions of data, for example attempting to develop prognostic algorithms from a particular set of genetic biomarkers. Such approaches, when touted as offering "complete" and accurate prediction models, serve to reinforce the implicit theoretical assumptions that inform the data acquisition—for example, the reductionist view that genes are the most important factors in disease prognosis, which at its most extreme verges on genetic determinism (Kohane, Masys, and Altman 2006). This is not to deny the importance of genetics in human disease, but simply to point out that there is no such thing as disinterested data analysis—all research occurs from a particular epistemic vantage point.

The theory-ladenness of observation, a critical insight of contemporary philosophy, has eroded the sharp distinction between epistemology and ontology, blurring the boundary between the knower and what is known (Marsonet 2018). The recognition that all experience is mediated by a conceptual scheme that shapes our knowledge of the world has far-reaching implications for the big data movement and clinical medicine in general (Quine 1980). These implications move beyond issues of epistemic interests and data selection to impact the medical ontologies constructed by our methods of measurement.

Measurement and Big Data Ontologies

The focus on measurement and quantification in medicine dates back to the Parisian empiricist school in the 19th century, which pioneered early methods of clinical measurement and medical statistics (Gavarret 1840; Matthews 1995). This quest for quantification took on new fervor with the rise of clinical epidemiology and evidence-based medicine in the 1990s, which some have argued can be understood as a shift not only towards reliance on "evidence" but also on clinical measurement (Bluhm and Borgerson 2011; McClimans 2013). The big data movement represents the latest trend in quantitative approaches to clinical medicine (Chin-Yee and Upshur 2018).

As with observation, measurement is also theory-laden (Smart 2017). Measurement proceeds from particular epistemic interests and is laden with theoretical assumptions, from the semantic understanding of the terms and values measured through to the technical apparatuses used in measurement (Kuhn 1962). Measurement, however, adds an additional layer to this problem. By representing concepts in concrete, quantitative terms, measurement constructs ontologies that can often conceal the theoretical assumptions involved in their creation. Without a critical lens, these ontologies can be taken as "givens," and misapprehended as unproblematic representations of "what there is"—or, often, "all there is."

Bradburn, Cartwright, and Fuller (2017) propose a three-stage theory of measurement for medicine that involves: (1) *characterization*, or defining the concept and its boundaries; (2) *representation*, or defining a metrical system to represent the concept; and (3) *procedures*, or establishing rules for applying metrical systems to tokens in measurement. Using this framework, we can diagnose several places where measurement and the resulting big data ontologies can go awry.

Bradburn and colleagues point out that many of the concepts used in healthcare research are *Ballung* concepts, which refers to concepts that can take on multiple meanings, with tokens being connected by "family-resemblance" rather than precise boundaries. Faced with challenges studying *Ballung* concepts, researchers often develop operationalized definitions of concepts to render *Ballung* concepts into "pinpoint" ones, with necessary and sufficient conditions. While operationalization might serve specific purposes—for example, to develop an objective and consensus definition for a particular research study—this move is also deflationary and eliminates any possibility for a "thick" description of a concept. Furthermore, as Bradburn and colleagues note, "operationalization makes knowledge accumulation difficult" (77). Using operationalized measures beyond their intended function raises a host of issues that pose particular problems for the big data movement.

Operationalized measures of concepts ranging from cardiovascular risk to cognitive function to performance status, among others, have proliferated in modern medicine, many of which are used beyond the purposes for which they were initially validated. Such measures have become part and parcel of day-to-day

clinical practice—for example, we routinely use the Glasgow Coma Scale (GCS) to indicate neurological status, or the Montreal Cognitive Assessment (MoCA) score to signify level of cognitive impairment (Nasreddine et al. 2005; Teasdale and Jennett 1974). These scores have been validated in particular cohorts, and in the right contexts can offer useful tools to aid clinical decision-making. However, if used outside of this context, these measures can be misleading. For example, using the MoCA to assess cognitive function in elderly inpatients with acute medical issues can result in false conclusions and misdiagnosis (Brown 2015). Such measures are often applied to label patients—for example, "The patient *is* MoCA 20" or "She *is* GCS 8"—conferring a misplaced concreteness within a narrow ontology of quantitative clinical scores.

These problems are not limited to clinical scores and classifications but extend to many laboratory measurements that are also validated in particular contexts. For example, the fecal occult blood test is a population-level screening test for colorectal cancer but is commonly (mis)applied to "diagnose" gastrointestinal bleeding in hospitalized patients (Sharma et al. 2001); the D-dimer assay is validated for evaluation of venous thromboembolism in patients with low pretest probability, although it is frequently (mis)used outside of this context (Smith et al. 2008).

Combining multiple operationalized measures, removed from context, as inputs into machine learning algorithms lands us in uncharted territory, and it is uncertain how resulting predictions will obtain beyond the initial study populations. In this way, big data and machine learning create new levels of abstraction, further distancing measurements from the concepts that they aim to characterize and represent (Winther 2014). This additional level of abstraction carries attendant issues of how to appropriately extrapolate and apply information to impact individual patient care, problems which have been discussed extensively in debates over the use of "evidence" in the clinical setting (Upshur 2005).

Upshur (2017) has raised this problem of using "at hand" data, such as data collected for purposes of insurance or clinical care, outside of the intended context. This issue is amplified by big data analytics. Additionally, many existing measures and classification systems, such as the ICD-10, fail to capture important clinical phenomena such as multiple concurrent chronic diseases and social determinants of health. Relying on narrow, "disease-centered" ontologies as the basis for training our machine learning algorithms risks leading us astray from our aim of improving clinical care. Rather than simply endorsing available data sources—seeking to expand databases of what is currently collated in electronic health records—we need to rethink the desiderata of measurement.

Criticisms of big data have largely focused on problems of data validity, citing poor-quality data as undermining knowledge produced by these approaches (Coveney, Dougherty, and Highfield 2016; Saracci 2018). And indeed, readily accessible data sourced from electronic health records or administrative registers

are often fraught with errors and omissions that limit their usefulness. However, as the above discussion highlights, the problem of data validity only scratches the surface, beneath which more fundamental epistemological and ontological issues loom, arising from the theory-ladenness of observation and the limitations of clinical measurement.

Furthermore, it is unclear how new conceptualizations, such as the integration of new findings at the bedside, can be integrated into AI approaches. The current framing of the utility of big data and AI in medicine is asymmetrically focused on novel knowledge generation arising from the machine, and not on the sentient and observant clinician. Increasing reliance on machine learning may eclipse the cultivation of observational skills and diminishing the sensibility of clinicians. Claims that diminish the importance of clinical acumen and defer to the superiority of AI should be based on more than just a handful of studies that focus on diagnosis reliant on image recognition: a set of well-designed fair comparisons across the wide spectrum of clinical reasoning tasks should be conducted. As with any new technology, systematic unbiased evaluation is required before widespread adaptation and uptake.

The Epistemological-Logical Problem

A second important problem that requires serious scrutiny is related to logic and justification. Some accounts of what is possible with machine learning and AI seem to suggest that AI somehow transcends logic and obviates the need for justification and explanation (Hinton 2018; Stead 2018). It is unclear how neural networks or other machine learning algorithms can be unrelated to some form of logic or rely upon a programming language that is not in some way dependent on logical operators. Given this, the constraints that logic itself faces will be applicable to AI. While these constraints in no way completely undermine the utility of AI, they do raise cautions about the limits of inference from knowledge produced by AI approaches. In this section we discuss the logical limits of algorithms in modelling clinical complexity and resulting problems with reliability and interpretability of AI approaches, which hinders their ability to provide clinical justification and explanation.

The Limits of Algorithms

Any complete account of the utility of AI and machine learning in medicine must explain how a finite set of algorithms, or bits of programming language, can map onto the complex realities of the phenomenal and biological world. Since algorithms are simply the specification of a set of rules to be followed in a programming language, there would need to be a complete identification of all processes—biological, social, psychological, historical—in that programming language isomorphic to the events. As is evident, this is a very complex and daunting undertaking. It is, in fact, impossible.

To start with, such events occur outside the programming language and can only be translated into the language once they have occurred, forming the dataset for an algorithm to access. This is a major non sequitur, given that past empirical success of a program does not guarantee its future success, in particular when dealing with the dynamics of complex systems.

Part of the impossibility is the incomplete nature of current empirical accounts of the phenomenon to be explained. Given that the universe is an open system with the laws of thermodynamics indicating high degrees of entropy, and that biological systems are subject to complex evolutionary dynamics that are incompletely understood and lack an agreed upon teleology for both cosmological and evolutionary forces, it is impossible for an algorithm to bootstrap its way out of these forces to make predictions.

Any attempt to construct a computer program that can predict a clinical outcome "with perfect accuracy" (as some proponents of big data analytics have suggested; Sniderman, D'Agostino, and Pencina 2015), would require a complete account of the phenomenon, which remains elusive in clinical medicine. Consider a common condition such as an asthma exacerbation, an event that some proponents have argued might be predicted by machine learning algorithms (Topol, Steinhubl, and Torkamani 2015). Growing research has focused on using AI to predict events and outcomes in asthma (Finkelstein and Jeong 2016). However, attempts to construct algorithms to predict such highly complex, contingent outcomes can encounter significant limitations. In some cases, they may result in counterintuitive, and even potentially dangerous results, such as the suggestion by one machine learning prognostic model that asthma was a protective factor in patients presenting to hospital with pneumonia (Cabitza, Rasoini, and Gensini 2017; Caruana et al. 2015).[2] Such errors occur because these programs remain incomplete and underdetermine the complexity of the systems they attempt to model. Even with an expanded set of explanatory variables, it is unlikely that such models would be able to account for rare or idiosyncratic factors, such as thunderstorms and other environmental disturbances that are known to have a significant impact in asthma epidemics (Cockcroft 2018). Given these limitations, a more modest perspective is necessary, far short of any notion of perfect prediction.

Interpretability, Reliability, and Explanation

Statistical approaches that are well accepted in the biomedical sciences have established traditions that relate the mathematical approaches used to fundamental principles. Even if most clinicians are not familiar with advanced statistics, there is a general consensus that such approaches should be reliable, valid, unbiased,

[2]In this particular study of patients presenting with pneumonia, history of asthma predicted an almost 50% reduction in mortality, which may be attributable to early admission to intensive care units, a factor which could not be coded in the machine learning algorithm (Caruana et al. 2015).

and subject to accountable scrutiny. In other words, results can be compared to some external standards, and critical appraisal frameworks exist to aid clinicians in determining how much confidence can be placed in such measures. Modern science is premised on the ability to critically scrutinize and evaluate measures, and in principle, results should be capable of independent replication. Such is not the case with AI approaches. Currently, few tools for the critical appraisal of tools derived from AI exist. No standards have been set to certify to clinicians that apps or algorithms meet basic standards of reliability and validity.

AI and machine learning are simply more complex varieties of models, and as such require fitting and testing. The empirical inputs in medicine must come from biological science, clinical observation, or patient history, and these must be integrated in order to make clinical sense. Since these evolve dynamically, they must be continuously integrated into programs. This raises important challenges for the reliability of predictive models in the clinical context, many of which are articulated by Sculley and colleagues (2014). Writing generally about challenges of training models and ensuring reliability of results, they discuss the CACE principle (Changing Anything Changes Everything). As they write:

> To make this concrete, imagine we have a system that uses features $x_1, \ldots x_n$ in a model. If we change the input distribution of values in x_1, the importance, weights, or use of the remaining $n - 1$ features may all change—this is true whether the model is retrained fully in a batch style or allowed to adapt in an online fashion. Adding a new feature x_{n+1} can cause similar changes, as can removing any feature x_j. No inputs are ever really independent. We refer to this here as the CACE principle: Changing Anything Changes Everything. The net result of such changes is that prediction behavior may alter, either subtly or dramatically, on various slices of the distribution. The same principle applies to hyper-parameters. Changes in regularization strength, learning settings, sampling methods in training, convergence thresholds, and essentially every other possible tweak can have similarly wide ranging effects. (2)

Similarly, Hinton (2018) argues that if the same neural network "is refit to the same data, but with changes in the initial random values of the weights, there will be different features in the intermediate layers" (1102). And further, "a neural net has many different and equally good ways of modelling the same data set" (1102). How we know that these are "equally good," however, is not something which is open to scrutiny.

If we are to effectively utilize machine learning and AI in medicine, these issues should give us pause for reflection. Extreme variation in model performance may make results unstable and unreliable. As well, and linked to the abovementioned issues of underdetermination, it is possible for multiple, potentially conflicting predictions to arise from any adjustment to the model. This clearly has significant implications for quality and safety of care based on such models. At a

minimum, it is something that warrants awareness among clinical communities to avoid the seductive convenience of using apps in practice without certification and validation of performance in a wide range of clinical contexts.

The ability to explain and offer justification for clinical decisions is also clearly central to clinical medicine. The need to provide reasons for clinical decisions, or to ground interpretations of a patient's illness, is a core ethical and epistemic responsibility of clinicians. Reliance on authority, inclination, and experience alone has become insufficient in an era of evidence-based medicine. Yet with the rise of big data and AI, there seems to be a dismissal of the need for such external reason giving in using the results generated by AI approaches.

We do not have a complete explanatory account of how the human brain processes information, creates and stores information, and makes inferences from a variety of inputs. This is a major goal of modern neuroscience. Many models are derived from our understanding of simpler model organisms, such as nematodes—organisms that themselves are still not completely understood. AI and neural networks, in particular, seek eventually to model human intelligence. Even though human brain function is orders of magnitude more complex, we nevertheless ask for external justification and public accountability from reasoning by human agents.

However, proponents of AI hold that such external scrutiny and reason giving accounts cannot be provided by neural networks. As Hinton (2018) writes:

> Understandably, clinicians, scientists, patients and regulators would all prefer to have a simple explanation for how a neural net arrives at its classification of a particular case. In the example of predicting whether a patient has a disease, they would like to know the hidden factors the network is using. However, when a deep neural network is trained to make predictions on a big data set, it typically uses layers of learned, non-linear features to model a huge number of complicated but weak regularities in data. It is generally infeasible to interpret these features because their meaning depends on complex interactions with uninterpreted features in other layers. (1102)

It is not that clinicians have a preference that stimulates the need for an explanation, but rather they have an ethical and epistemic obligation to do so to the best of their ability. We do not permit such shoulder shrugging for human decisions: we require reasons and a human agent to assume responsibility. Deference to the inexplicable is not sound policy.

A central aim of AI and machine learning in medicine is to offer personalized solutions by leveraging large amounts of data from a range of sources. Included in this aim is the goal of improving the "patient experience" and helping to support more "person-centered" health care. This goal, however, may be stifled on several fronts. As demonstrated in the first section, data included in predictive algorithms are shaped by epistemic interests that often diverge from those

of patients, and clinical measures remain rooted in narrow, "disease-centered" ontologies. As this section has demonstrated, current computer programs face significant limitations, particularly in attempts to model the dynamics of complex systems and respond to the subtle changes in human variables that are part and parcel of clinical medicine. If AI is to improve patient experiences, then we must design systems that are responsive to those experiences, and that can consider "person-centered" data.

To what extent can this data be captured and purposed for machine learning algorithms? This question leads us to the third problem, the phenomenological problem, which will be the focus of the following section.

The Phenomenological Problem

Even if big data and machine learning were able to successfully utilize broader datasets to construct complex predictive algorithms, these approaches are still unlikely to capture more nuanced variables, in particular those concerning patient values and preferences, which are difficult to code as inputs into quantitative models. In the same way, much of medicine's rich and informative qualitative data are currently systematically excluded from big data approaches (Chin-Yee and Upshur 2018; Upshur, VanDenKerkhof, and Goel 2001).

Some proponents of AI in medicine are sensitive to this issue and have ambitions to utilize "person-centered" variables in the development of these technologies (Topol 2016). However, these efforts may also be misguided. The epistemic vantage point of big data and machine learning remains situated within a third-person perspective, and therefore encounters inherent limitations in capturing the lived experience of patients. We explore this issue below, drawing on work in the phenomenology of illness to show how patients' lived experiences constitute a central, nontrivial element of clinical medicine that cannot be ignored in health-care research and clinical reasoning. Phenomenology is also essential to understanding clinical judgment, and how physicians interpret patients' first-person experiences to guide action in a particular case. In what follows we discuss how the inability to account for the phenomenological perspectives of both clinicians and patients represents a core limitation of big data and AI in medicine.

The Phenomenology of Illness

The phenomenology of illness has become an influential movement in the philosophy of medicine, with important applications in clinical practice and medical education. This movement focuses on the first-person experience of illness, in contrast to the third-person perspective offered by the naturalistic biomedical model. The phenomenology of illness is often framed in opposition to biomedicine's view of disease as biological dysfunction, which phenomenologists argue

objectifies patients and devalues subjective illness experiences. Leading contemporary authors of this movement include S. Kay Toombs (1988), Havi Carel (2016), and Fredrik Svenaeus (2013), among others, who draw on work by 20th-century phenomenologists from Edmund Husserl, Martin Heidegger, and Hans-Georg Gadamer to Jean-Paul Sartre and Maurice Merleau-Ponty. These authors offer compelling arguments for the value of a phenomenological approach to move beyond biomedicine's third-person generalizations and illuminate the first-person experience of illness while attending to its diversity and complexity.

Embodiment is a key concept in the phenomenology of illness. Bodily perception is imbued with meaning, which is prior to any isolated "sense data" (Toombs 1988). Contrary to the logical positivist view, one does not perceive the world as isolated sense-data and then proceed to construct observation statements with empirical content and meaning. Rather, meaning is antecedent—there is an "immanent significance" that discloses itself to the body in perception (Merleau-Ponty 1945, 26).

Phenomenology highlights how meaning and context are basic features of experience that are not accounted for by logical positivism's impoverished sense-data. This approach also shows why the first-person experience in health care cannot be captured by simply considering context-free statements of individual preferences and values. Such attempts remain tied to atomistic notions of data inputs, which forgo nuance and complexity by failing to situate perception and experience within a broader life-world.

The first-person perspective afforded by phenomenology's attention to the lived body helps us to better understand the experience of illness. As Toombs (1988) points out:

> Arthritis represents not so much an inflammation of the joints as it does the "inability to" button my shirt, swing a golf club, play tennis. In illness bodily intentionality is frustrated . . . . For the person with angina, for example, stairs which in health were simply there "to be climbed," are now obstacles "to be circumvented," "avoided," or even "feared." (208)

Such perspectives offer valuable insights to practitioners and may help realign care to best serve patients' needs. Phenomenological approaches have informed the creation of patient resources and education tools for health-care professionals (Carel 2010, 2012). Carel (2016) proposes a "phenomenological toolkit" to enable patients to better articulate their experiences and to allow clinicians to hone the "epistemic sensibilities and skills" (183). A key step in the phenomenological toolkit is "bracketing the natural attitude," which entails momentarily setting aside narrow biomedical ontologies in order to foster a richer, more comprehensive understanding of how illness impacts a person's being in the world (200).

Carel and Kidd (2014) argue for a phenomenological approach as a means of countering epistemic injustices that arise in health care and can result in the

dismissal or disbelief of patient experiences (Fricker 2007). Big data and machine learning's neglect of the phenomenology of illness risks exacerbating epistemic injustice: not only is first-person experience excluded from these quantitative tools, but this knowledge is robbed of clinical import in the face of powerful algorithms fueled by swaths of data. This echoes criticisms raised a decade ago against evidence-based medicine hierarchies, which by privileging particular research methodologies and definitions of evidence devalued other sources of knowledge, such as patient and physician testimonies (Bluhm and Borgerson 2011; Goldenberg 2006).

We have argued elsewhere that engagement with first-person experience is essential to clinical judgment (Chin-Yee and Upshur 2015, 2018). In addition to phenomenology, other methods have been proposed to support this engagement, such as narrative or historical approaches (Charon 2006; Chin-Yee and Upshur 2015). There are similarities and differences between these approaches, as there are distinctions that can be drawn between medical phenomenologists. Nonetheless, in our view the phenomenology of illness represents a particularly well-articulated position emerging from a strong philosophical framework that offers a range of clinically useful applications. The phenomenological approach also best highlights the deficiencies of the epistemic attitude engendered by the big data and machine learning movement in medicine.

Phenomenology and AI

Phenomenology's bracketing of the natural attitude is critical to opening up a clinician's interpretive horizon and transcending the limits of narrow "disease-centered" ontologies (Chin-Yee, Messinger, and Young 2019; Gadamer 1996). This step is necessary for the provision of individualized, compassionate care that is sensitive to context. As we saw in the first section, big data ontologies constructed from "at-hand" data encounter a range of limitations, including their potential to misguide clinical care and research when extrapolating beyond domains where such ontologies may apply.

By excluding the phenomenological perspective, big data and machine learning leave out a crucial component of clinical reasoning. The first-person knowledge of both patients and physicians features center stage in the clinical encounter. For the clinician, this perspective is what enables identification of the most salient data and resources to allow for diagnosis and management in each individual case. This ability to recognize salience and adapt action to a particular context is a unique feature of human intelligence that has been recognized at least since Aristotle, and is captured by the notion of practical knowledge, or *phronesis*. Phronesis, which Kathryn Montgomery (2005) defines as the "flexible, interpretive capacity that enables moral reasoners to determine the best action to take when knowledge depends on the circumstance" (5), has been identified by several authors as a defining trait of a good physician (Pellegrino and Thomasma 1981; Svenaeus 2003).

Programming computers to exercise practical reasoning and demonstrate "commonsense" understanding has proved a challenge for researchers in AI. According to some critics, this failure is the direct result of the inability of computers to incorporate a phenomenological perspective. In 1972, Heideggerian scholar Herbert Dreyfus published an influential manifesto entitled *What Computers Can't Do: A Critique of Artificial Reason,* in which he argued that the limits encountered by AI were a symptom of the Western philosophy's misapprehension of human reason. Reissued in 1992 as *What Computers Still Can't Do*, Dreyfus's work traces this misunderstanding back to Plato's assertion that "all knowledge must be stateable in explicit definitions" (67), through to Descartes's position that "all understanding consists in forming and using appropriate symbolic representations . . . built up out of primitive ideas or elements" (xi). This view of human reason is elaborated by Kant's notion that "all concepts are rules for relating such elements," and formalized by Frege so that concepts could be manipulated "without intuition or interpretation" (xi). Dreyfus saw in logical positivism the convergence of rationalist and empiricists traditions, establishing a worldview consisting in discrete facts or "logical atoms" from which the mind constructs symbolic representations that are manipulated by formal rules and algorithms. For Dreyfus, contemporary AI represented the fullest expression of logical positivism and the culmination of its worldview—he also saw it as a "degenerating" research program, and its failures as evidence of the limitations of this philosophy (Lakatos 1975).

We have seen previously how logical positivism has been seriously challenged by post-positivist philosophies of science. Dreyfus draws on the phenomenology of Heidegger and Merleau-Ponty to argue that analogies between the human mind and computers fundamentally misrepresent the nature of experience and being-in-the-world. According to the phenomenologist, what we perceive is not raw data but "immanent significance," shaped by context and directed towards action (Merleau-Ponty 1945, 26). Just as this perspective illuminates the experience of illness, it also highlights why clinical judgment cannot be reduced to fact-gathering and the application of rules. In each clinical encounter, the clinician begins by attempting to understand the purpose of the consultation—what Gadamer (1975) might call "the question." It is via "the question," which arises from a concrete situation, that all "facts" and "evidence" are given meaning and interpreted to guide action. As Gadamer put it, "the path to all knowledge leads through the question" (371). This question may be implied, but it is often not immediately apparent and only emerges and evolves through a process of dialogue and careful interpretation.

Phenomenology's recognition of the interplay between fact and situation, whereby data derives its meaning from context, and context continually changes, touches on a central challenge for AI. According to Dreyfus, for AI to have any hope of simulating human intelligence, a phenomenological approach that

best describes how humans engage with the world is needed. Forty years since its initial publication, Dreyfus's critique remains highly relevant, in particular for considering the prospects of AI in medicine. Although his argument was primarily directed at what has been termed "Good Old-Fashioned AI," the engineered, rule-based systems that dominated AI research at the time of his writing, many of the current medical applications of AI bear strong resemblance to these technologies. In subsequent editions, Dreyfus did address newer machine learning technologies, such as artificial neural networks, the predecessor of current "deep learning" algorithms, which are now being widely applied in health care (Hinton 2018). While these systems have less reliance on preprogrammed rules, they still suffer from an inability to adapt to context and condition on the most relevant features within a given situation—unlike human intelligence, they cannot cope with structural change in their environment. Big data does not overcome this problem. As Dreyfus (1992) puts it, "expert know-how cannot be put into the computer by adding more facts, since the issue is which is the current correct perspective from which to determine which facts are relevant" (xlii).

We are still some ways from developing AI that reproduces human reasoning and practical knowledge, of which clinical judgment is a paradigmatic example. However, as we have argued, in its exclusion of the phenomenological perspective, the growth of big data and machine learning may have more immediate consequences in clinical medicine. These range from exacerbating epistemic injustice and devaluing the first-person knowledge of patients to mischaracterizing clinical judgment and undermining the interpretive, dialogic aspects of the patient-clinician relationship. By elevating these technologies as exemplars of clinical reasoning, rather than seeing them for what they are—potentially useful adjuncts within restricted domains—we risk leaving out fundamental features of the clinical encounter that these approaches exclude.

Conclusion

The rise of big data and AI in health care raises significant philosophical issues with pressing relevance to clinical medicine. We have highlighted three philosophical challenges faced by this movement: the epistemological-ontological problem arising from the theory-ladenness of data and measurement; the epistemological-logical problem surrounding the inherent limitations of algorithms and issues with reliability and interpretability; and the phenomenological problem concerning exclusion of first-person experience of clinicians and patients. Some may worry that our analysis undermines claims to the objectivity of science and leads to an indefensible strain of relativism. We reject that conclusion. Our analysis is complementary with a pluralist account of medical science which, although anti-foundationalist, is compatible with affirming the importance of science in medical progress (Upshur 2002).

Claims that physicians will soon be replaced by AI are indeed overstated (Naylor 2018); however, the philosophical issues highlighted here suggest more proximate detrimental impacts of these technologies, ranging from the privileging of quantitative data and the exclusion of first-person knowledge to the underdetermination of clinical complexity by AI algorithms. While we do not expect argumentation to slow down the current enthusiasm for applying these technologies in health care, we do hope that our arguments will contribute to the evolution of critical appraisal standards to evaluate the impact of these technologies as they enter clinical care.

References

Amato, F., et al. 2013. "Artificial Neural Networks in Medical Diagnosis." *J Appl Biomed* 11 (2): 47–58.

Anderson, C. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired*, June 23. https://www.wired.com/2008/06/pb-theory/.

Beam, A. L., and I. S. Kohane. 2018. "Big Data and Machine Learning in Health Care." *JAMA* 319 (13): 1317–18.

Bennett, C. C., and K. Hauser. 2013. "Artificial Intelligence Framework for Simulating Clinical Decision-Making: A Markov Decision Process Approach." *Artif Intell Med* 57 (1): 9–19.

Bluhm, R., and K. Borgerson. 2011. "Evidence-Based Medicine." In *Handbook of the Philosophy of Medicine,* ed. D. M. Gabbay, P. Thagard, and J. Woods, 1–36. Oxford: North Holland.

Bradburn, N. M., N. L. Cartwright, and J. Fuller. 2017. "A Theory of Measurement." In *Measurement in Medicine: Philosophical Essays on Assessment and Evaluation*, ed. L. McClimans, 73–87. London: Rowman and Littlefield.

Brown, J. 2015. "The Use and Misuse of Short Cognitive Tests in the Diagnosis of Dementia." *J Neurol Neurosurg Psychiatry* 86 (6): 680–85.

Cabitza, F., R. Rasoini, and G. F. Gensini. 2017. "Unintended Consequences of Machine Learning in Medicine." *JAMA* 318 (6): 517–18.

Carel, H. 2010. "Phenomenology and Its Application in Medicine." *Theor Med Bioeth* 32 (1): 33–46. DOI: 10.1007/s11017-010-9161-x.

Carel, H. 2012. "Phenomenology as a Resource for Patients." *J Med Philos* 37 (2): 96–113.

Carel, H. 2016. *Phenomenology of Illness*. Oxford: Oxford University Press.

Carel, H., and I. J. Kidd. 2014. "Epistemic Injustice in Healthcare: A Philosophical Analysis." *Med Health Care Philos* 17 (4): 529–40.

Carnap, R. 1919. *Logical Syntax of Language*. London: Routledge and Kegan Paul.

Caruana, R., et al. 2015. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–30. New York: Association for Computing Machinery.

Chaitin, G. 2002. "Computers, Paradoxes and the Foundations of Mathematics." *Am Scientist* 90 (2): 164–71.

Chaitin, G. J. 2004. *Algorithmic Information Theory*. Cambridge: Cambridge University Press.

Charon, R. 2006. *Narrative Medicine: Honoring the Stories of Illness*. Oxford: Oxford University Press.

Chin-Yee, B. H. 2014. "Underdetermination in Evidence-Based Medicine." *J Eval Clin Pract* 20 (6): 921–27.

Chin-Yee, B. H., and R. E. G. Upshur. 2015. "Historical Thinking in Clinical Medicine: Lessons from R. G. Collingwood's Philosophy of History." *J Eval Clin Pract* 21 (3): 448–54.

Chin-Yee, B., and R. Upshur. 2018. "Clinical Judgement in the Era of Big Data and Predictive Analytics." *J Eval Clin Pract* 24 (3): 638–45.

Chin-Yee, B., A. Messinger, and L. T. Young. 2019. "Three Visions of Doctoring: A Gadamerian Dialogue." *Adv Health Sci Educ* 24: 403. DOI: 10.1007/s10459-018-9824-3.

Cockcroft, D. W. 2018. "Epidemic Thunderstorm Asthma." *Lancet Planet Health* 2 (6): e236–e237.

Coveney, P. V., E. R. Dougherty, and R. R. Highfield. 2016. "Big Data Need Big Theory Too." *Philos Trans A Math Phys Eng Sci* 374. DOI: 10.1098/rsta.2016.0153.

Dreyfus, H. L. 1992. *What Computers Still Can't Do*. Cambridge: MIT Press.

Esteva, A., et al. 2017. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature* 542 (7639): 115–18.

Finkelstein, J., and I. C. Jeong. 2016. "Machine Learning Approaches to Personalize Early Prediction of Asthma Exacerbations." *Ann NY Acad Sci* 1387 (1): 153–65.

Fleming, N. 2018. "How Artificial Intelligence Is Changing Drug Discovery." *Nature* 557 (7707): S55–S57.

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Gadamer, H.-G. 1975. *Truth and Method*. London: Bloomsbury, 2013.

Gadamer, H.-G. 1996. *The Enigma of Health*. Redwood City: Stanford University Press.

Gavarret, J. 1840. *Principes généraux de statistique médicale*. Paris: Bechet Jeune et Labé.

Goldenberg, M. J. 2006. "On Evidence and Evidence-Based Medicine: Lessons from the Philosophy of Science." *Soc Sci Med* 62 (11): 2621–32.

Goldenberg, M. J. 2009. "Iconoclast or Creed? Objectivism, Pragmatism, and the Hierarchy of Evidence." *Perspect Biol Med* 52 (2): 168–87.

Goodman, K. 2016. *Ethics, Medicine, and Information Technology: Intelligent Machines and the Transformation of Health Care*. Cambridge: Cambridge University Press.

Gulshan, V., et al. 2016. "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs." *JAMA* 316 (22): 2402–9.

Hinton, G. 2018. "Deep Learning: A Technology with the Potential to Transform Health Care." *JAMA* 320 (11): 1101–2.

Khoury, M. J., and S. Galea. 2016. "Will Precision Medicine Improve Population Health?" *JAMA* 316 (13): 1357–58.

Kohane, I. S., D. R. Masys, and R. B. Altman. 2006. "The Incidentalome: A Threat to Genomic Medicine." *JAMA* 296 (2): 212–15.

Kuhn, T S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 2012.

Lakatos, I. 1975. "Falsification and the Methodology of Scientific Research Programmes." In *Can Theories Be Refuted? Essays on the Duhem-Quine Thesis*, ed. S. Harding, 205–59. Dordrecht: D. Reidel.

Longino, H. E. 2002. *The Fate of Knowledge*. Princeton: Princeton University Press.

Livingstone, S. G., et al. 2015. "Much Ado About Omics: Welcome to 'the Permutome.'" *J Eval Clin Pract* 21 (6): 1018–21.

Marsonet, M. 2018. "On the Ontology/Epistemology Distinction" In *The Map and the Territory: Exploring the Foundations of Science, Thought and Reality*, ed. S. Wuppuluri and F. A. Doria, 15–34. Cham: Springer.

Matthews, J. R. 1995. *Quantification and the Quest for Medical Certainty*. Princeton: Princeton University Press.

McClimans, L. 2013. "The Role of Measurement in Establishing Evidence." *J Med Philos* 38 (5): 520–38.

Merleau-Ponty, M. 1945. *Phenomenology of Perception*. London: Routledge, 2013.

Montgomery, K. 2005. *How Doctors Think: Clinical Judgment and the Practice of Medicine*. Oxford: Oxford University Press.

Nasreddine, Z. S., et al. 2005. "The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool for Mild Cognitive Impairment." *J Am Geriatr Soc* 53(4): 695–99.

Naylor, C. D. 2018. "On the Prospects for a (Deep) Learning Health Care System." *JAMA* 320 (11): 1099–1100.

Obermeyer, Z., and E. J. Emanuel. 2016. "Predicting the Future: Big Data, Machine Learning, and Clinical Medicine." *N Engl J Med* 375 (13): 1216–19.

Obermeyer, Z., and T. H. Lee. 2017. "Lost in Thought: The Limits of the Human Mind and the Future of Medicine." *N Engl J Med* 377 (13): 1209–11.

Ow, G. S., and V. A. Kuznetsov. 2016. "Big Genomics and Clinical Data Analytics Strategies for Precision Cancer Prognosis." *Sci Rep* 6 (36493): 1–13.

Pearl J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Pellegrino, E. D., and D. C. Thomasma. 1981. *A Philosophical Basis of Medical Practice*. Oxford: Oxford University Press.

Popper, K. 1962. *Conjectures and Refutations*. London: Routledge, 2014.

Quine, W. V. O. 1980. *From a Logical Point of View: Nine Logico-Philosophical Essays*. Cambridge: Harvard University Press.

Rumsfeld, J. S., K. E. Joynt, and T. M. Maddox. 2016. "Big Data Analytics to Improve Cardiovascular Care: Promise and Challenges." *Nat Rev Cardiol* 13 (6): 350–59.

Saracci, R. 2018. "Epidemiology in Wonderland: Big Data and Precision Medicine." *Eur J Epidemiol* 33 (3): 245–57.

Sculley, D. 2014. "Machine Learning: The High-Interest Credit Card of Technical Debt." *SE4ML* 1–9.

Sharma, V. K., et al. 2001. "An Audit of the Utility of In-Patient Fecal Occult Blood Testing." *Am J Gastroenterol* 96 (4): 1256–60.

Smart, B. 2017. "How Evidence-Based Medicine Highlights Connections Between Measurement and Evidence." In *Measurement in Medicine: Philosophical Essays on Assessment and Evaluation*, ed. L. McClimans, 3–10. London: Rowman and Littlefield.

Smith, C., et al. 2008. "Is Pretest Probability Assessment on Emergency Department Patients with Suspected Venous Thromboembolism Documented Before SimpliRED D-Dimer Testing?" *CJEM* 10 (6): 519–23.

Sniderman, A. D., R. B. D'Agostino, and M. J Pencina. 2015. "The Role of Physicians in the Era of Predictive Analytics." *JAMA* 314 (1): 25–6.

Stead, W. W. 2018. "Clinical Implications and Challenges of Artificial Intelligence and Deep Learning." *JAMA* 320 (11): 1107–8.

Svenaeus, F. 2003. "Hermeneutics of Medicine in the Wake of Gadamer: The Issue of Phronesis." *Theor Med Bioeth* 24 (5): 407–31.

Svenaeus, F. 2013. *The Hermeneutics of Medicine and the Phenomenology of Health*. Dordrecht: Springer.

Teasdale, G., and B. Jennett. 1974. "Assessment of Coma and Impaired Consciousness: A Practical Scale. *Lancet* 2 (7872): 81–84.

Thornton, J. 2015. "What You Need to Know to Make the Most of Big Data in Biology." *Lancet* 385: S5–S6.

Toombs, S. K. 1988. "Illness and the Paradigm of Lived Body." *Theor Med* 9 (2): 201–26.

Topol, E. 2016. *The Patient Will See You Now*. New York: Basic Books.

Topol, E. J., S. R. Steinhubl, and A. Torkamani. 2015. "Digital Medical Tools and Sensors." *JAMA* 313 (4): 353–54.

Tu, J. V. 1996. "Advantages and Disadvantages of Using Artificial Neural Networks Versus Logistic Regression for Predicting Medical Outcomes." *J Clin Epidemiol* 49 (11): 1225–31.

Upshur, R. E. G. 2002. "If Not Evidence, Then What? Or Does Medicine Really Need a Base?" *J Eval Clin Pract* 8 (2): 113–19.

Upshur, R. E. G. 2005. "Looking for Rules in a World of Exceptions: Reflections on Evidence-Based Practice." *Perspect Biol Med* 48 (4): 477–89.

Upshur, R. E. G. 2017. "Measurement, Multiple Concurrent Chronic Conditions, and Complexity." In *Measurement in Medicine: Philosophical Essays on Assessment and Evaluation*, ed. L. McClimans, 133–48. London: Rowman and Littlefield.

Upshur, R. E. G., E. G. VanDenKerkhof, and V. Goel. 2001. "Meaning and Measurement: An Inclusive Model of Evidence in Health Care." *J Eval Clin Pract* 7 (2): 91–96.

Verghese, A., N. H. Shah, and R. A. Harrington. 2018. "What This Computer Needs Is a Physician." *JAMA* 319 (1): 19–2.

Winther, R. G. 2014. "James and Dewey on Abstraction." *Pluralist* 9 (2): 1–28.